11:46:22 OCA PAD AMENDMENT - PROJECT HEADER INFORMATION 10/13/95

Active

Project #: E-24-625 Cost share #: Rev #: 6
Center # : 10/24-6-R7676-0A0 Center shr #: OCA file #:
Work type : RES
Contract#: F49620-93-1-0043 Mod #: BUDGET REVISION Document : GRANT
Prime #: Contract entity: GTRC

Subprojects ? : N CFDA: 12.800
Main project #: PE #: 61102F

Project unit: ISYE Unit code: 02.010.124
Project director(s):
ALEXOPOULOS C ISYE (404)894-2361

Sponsor/division names: AIR FORCE / BOLLING AFB, DC
Sponsor/division codes: 104 / 001

Award period: 921101 to 950930 (performance) 951130 (reports)

Sponsor amount	New this change	Total to date
Contract value	0.00	88,167.00
Funded	0.00	88,167.00
Cost sharing amount		0.00

Does subcontracting plan apply ?: N

Title: A CLASS OF METHODS FOR ANALYZING STOCHASTIC SYSTEMS

PROJECT ADMINISTRATION DATA

OCA contact: Anita D. Rowland 894-4820

Sponsor technical contact

DR. JON A. SJOGREN
(202)767-4940

AFOSR/NM
BUILDING 410
BOLLING AFB, DC 20332-6448

Sponsor issuing office

KAREN BUCK
(202)767-4943

AFOSR/PKA
BUILDING 410
BOLLING AFB, DC 20332-6448

Security class (U,C,S,TS) : U ONR resident rep. is ACO (Y/N): N
Defense priority rating : NA X supplemental sheet
Equipment title vests with: Sponsor GIT X
<$5000 TITLE=GIT AT TIME OF ACQUISITION. REF. PG 1 OF GRANT.
Administrative comments -
PROCESSED REQUEST DTD 10.1.95.
**PLEASE PROVIDE STATUS ON DELIVERABLES.

GEORGIA INSTITUTE OF TECHNOLOGY
OFFICE OF CONTRACT ADMINISTRATION

NOTICE OF PROJECT CLOSEOUT

Closeout Notice Date 12/11/95

Project No. E-24-625 Center No. 10/24-6-R7676-0A0

Project Director ALEXOPOULOS C School/Lab ISYE

Sponsor AIR FORCE/BOLLING AFB, DC

Contract/Grant No. F49620-93-1-0043 Contract Entity GTRC

Prime Contract No. _____

Title A CLASS OF METHODS FOR ANALYZING STOCHASTIC SYSTEMS

Effective Completion Date 950930 (Performance) 951130 (Reports)

Closeout Actions Required:	Y/N	Date Submitted
Final Invoice or Copy of Final Invoice	Y	
Final Report of Inventions and/or Subcontracts	Y	
Government Property Inventory & Related Certificate	N	
Classified Material Certificate	N	
Release and Assignment	N	
Other _____	N	

Comments _____

Subproject Under Main Project No. _____

Continues Project No. _____

Distribution Required:

Project Director	Y
Administrative Network Representative	Y
GTRI Accounting/Grants and Contracts	Y
Procurement/Supply Services	Y
Research Property Managment	Y
Research Security Services	N
Reports Coordinator (OCA)	Y
GTRC	Y
Project File	Y
Other _____	N
_____	N

NOTE: Final Patent Questionnaire sent to PDPI.

E-24-625
/

Annual Report
Grant F49620–93–1–0043
Air Force Office of Scientific Research

Christos Alexopoulos
School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332–0205

December 29, 1993

1 Introduction

This document is the first annual report for grant F49620–93–1–0043. My research resulted in five papers and two completed doctoral dissertations. Three of these papers have been accepted for publication. In addition, four other papers are in the final developmental stage and will be submitted for publication shortly. Furthermore, one doctoral student works under my supervision on problems that should be of great interest to Air Force laboratories and the airline industry.

Sections 2–5 review results from my papers and section 6 concludes with a brief description of my ongoing research.

2 Methods for Computing Network Performability Measures

Stochastic networks are used to model a variety of industrial and communications systems. These systems include data communications networks, voice communications networks, transportation networks. computer architectures, and electrical power systems. Stochastic networks are modeled by graphs in which each arc, and probably each node, is assigned a nonnegative random weight. The component weights have interpretations depending on the type of network under consideration. My research has focused on the evaluation of general *performability* measures and considered the following types of systems:

Flow Networks The nodes model distribution centers and the arcs represent the means of transmitting commodities between pairs. The nodes are classified into sources, demand nodes, and transshipment nodes. The weight on each arc and transshipment node represents a capacity that limits the total amount of commodity that can be transmitted. An arc may also be associated with a random cost per unit of transmitted commodity.

The following are typical measures of interest: (a) The probability that the demands can be satisfied; (b) The probability that a given set of links and nodes limits commodity transmission when the demands cannot be satisfied; (c) The expected amount of unsupplied flow when the demands cannot be satisfied; (d) The probability that the total cost for satisfying the demands does not exceed a specified value.

Transportation Networks The arcs represent sections of routes and the nodes represent intersections of routes. The weight of an arc represents its length or travel time. A list of interesting problems includes the computation of: (a) The distribution of the shortest path length from a source s to a destination t; (b) The probability that a given arc belongs to a shortest path.

Undirected Networks Networks with undirected arcs are often used for modeling communications systems or for solving a variety of problems. An example is a graph whose arcs have random costs and the objective is the evaluation of the probability that the nodes can be connected via a spanning tree whose total cost does not exceed a given budget. Another example is a bipartite graph where the left-hand set of nodes represents personnel and the right-hand set represents jobs. An arc (i, j) indicates that person i can perform job j at a (random) cost C_{ij}. The objective is the identification of a matching between personnel and jobs with minimum total expected cost.

The majority of problems for computing performability measures for stochastic networks are *#P-hard.* This property has motivated the research for approximation methods. One class of these methods attempts to compute bounds while another class focuses on Monte Carlo estimation methods.

My research has focused on a methodology that are based on iteration and, in short, evaluate a performability measure as follows: At each iteration, a subset of the system state space is partitioned into sets with known contribution to the measure, sets with zero contribution, and *undetermined* sets whose value is unknown. The method continues in the same fashion until no undetermined sets remain. The proposed methods have the following important properties:

- After each iteration, they produce lower and upper bounds that improve continuously.
- The bounds along with the remaining undetermined sets can be used for designing Monte Carlo sampling plans that (a) yield estimates with variance smaller by several orders of magnitude than the variance of the respective estimates produced by a crude Monte Carlo experiment with equal sample size and (b) take less time than the crude experiment.

Publication [2] makes two contributions: (a) It corrects several errors in well-known algorithms by Doulliez and Jamoulle [7] for computing performability measures for flow networks, and (b) disputes previously made claims on the applicability of state space partitioning methods. It should be mentioned that the algorithm of Doulliez and Jamoulle is the most frequently used approach for analyzing electrical power systems.

Technical report [3] proposes partitioning methods for computing measures related to shortest paths. The dissertation of my former student Jay Jacobson proved theoretical properties of these methods and examined their applicability on stochastic minimum spanning tree problems and minimum cost flow problems. Overall, this thesis made the following contributions:

- It proved that several stochastic network problems are *#P-hard* and presented efficient methods for solving these problems exactly. In particular, the *matroidal*

structure of the minimum spanning tree problem gave rise to an impressive algorithm for computing the probability that an arc belongs to a minimum spanning tree.

- It advanced the understanding of state space partitioning methods. In doing so, it made these methods more accessible and drew strong conclusions about the performance of certain types of partitions.
- It proposed areas in which further gains can be made with regards to these powerful computational techniques.

Two joint publications are in their final processing stage and will be submitted for publication during the next 2–3 months.

3 Stochastic Processes

The dissertation of El-Tannir, co-advised by myself and Richard Serfozo, studied two topics:

Multivariate Generalizations of Markov Modulated Processes Markov modulated processes model queueing systems where the arrival and service rates vary according to a Markov process independently of the number of customers in the system. These processes, however, do not cover systems where the arrival and service rates depend on the number of customers present. An example is an $M/M/Y$ system where the number of servers $Y(t)$ at time t is a Markov process with rates that depend on the number of customers present.

The paper by Alexopoulos, El-Tannir and Serfozo [4] studies a family of multivariate Markov processes where transitions can take place simultaneously and the rate at which a set of components changes state depends on the state of the remaining components. This family covers a wide range of Markov processes including Markov modulated processes, Markovian queues with variable capacity, and standard network processes such as closed Jackson network processes. The paper makes the following three contributions: (*a*) It identifies processes whose stationary distributions have product form; (*b*) It presents approximations for stationary distributions. Theorem 13 (p. 8) gives an approximation for a bivariate process (X, Y) that is based on an "auxiliary" process with averaged rates while section 3.2 generalizes this result for multivariate processes. When the component X has n states and the component Y has m states, the computation of the approximate distribution requires the solution of $m+1$ subsystems each with dimensions $n \times n$ instead of solving an $(mn) \times (mn)$ system. Additional approximations are described in sections 3.3 and 3.4 while section 4 contains several examples.

Starlike Markov Processes Our research addressed the following question: "Can the stationary distribution of a Markov process be obtained by pasting together several stationary distributions restricted to certain subspaces?" Eltannir's study described a class of "starlike" processes that have this cut-and-paste or divide-and-conquer property. The state space of a starlike process can be partitioned into a "central" set and a collection of "peripheral" sets, and the process cannot move from a peripheral set to another peripheral set unless it passes through the central set. We derived necessary and sufficient conditions for the stationary distribution of a starlike process to be expressed in terms of the stationary distributions of Markov processes restricted to the union of the central set and a peripheral set. A joint publication is in the final developmental stages and will be submitted for publication.

4 Distribution-Free Confidence Intervals

My work on distribution-free confidence intervals was motivated by problems related to the estimation of performability measures (paper [1]) or comparisons between alternative system designs (paper [5]).

Publication [1] developed confidence intervals for the ratio of two means whose estimation is based independent, identically distributed random pairs with bounded and ordered components. Emphasis is given to the case in which each component can be expressed as the product of a Bernoulli and a bounded random variable. The proposed intervals result from distribution-free, Bernstein-type bounds on error probabilities, are valid for every sample size, and their asymptotic width decreases at the same rate as that of confidence intervals based on the central limit theorem. Experimental results showed that the proposed intervals are conservative with superior coverage for small sample sizes (≤ 50). This superiority over "normal" confidence intervals makes them useful for experiments where replications are expensive.

Publication [5] proposed distribution-free confidence intervals for multinomial experiments. Below, I briefly discuss the single, but important, result of this paper. Let $p = (p_1, p_2, \ldots, p_k)$ denote the unknown cell probabilities and suppose that we draw n samples. Let $n = (n_1, n_2, \ldots, n_k)$ be the observed counts and denote the observed cell proportions by $\hat{p}_i = n_i/n, i = 1, \ldots, k$.

The proposed confidence intervals have the form

$$\hat{p}_i \pm t/\sqrt{n}, \qquad i = 1, \ldots, k$$

with simultaneous confidence coefficient

$$\Pi(k, p; n, t) = P\left[\bigcap_{i=1}^{k} |\hat{p}_i - p_i| < t/\sqrt{n}\right] \geq 1 - \alpha, \qquad \alpha \in (0, 1).$$

Our methodology is based on the bound

$$\Pi(k,p;n,t) \geq G(k,p;n,t) \geq \min_p G(k,p;n,t),$$

where

$$G(k,p;n,t) = 1 - 2\sum_{i=1}^{k} \exp\left\{-nt\left[\left(1 + \frac{p_i(1-p_i)\sqrt{n}}{t}\right)\ln\left(1 + \frac{p_i(1-p_i)\sqrt{n}}{t}\right) - 1\right]\right\},$$

and finds the smallest t such that

$$\min_p G(k,p;n,t) = 1 - \alpha.$$

In fact, a lengthy proof shows that $G(k,p;n,t)$ is minimized when $p_1 = \cdots = p_m = 1/m$ for some $2 \leq m \leq k$ and $p_i = 0$ for $i > m$.

The following table summarizes our findings. The last column lists asymptotically valid "normal" confidence intervals from Fitzpatrick and Scott [9]. For example, when $n = 500$ the intervals $\hat{p}_i \pm 1.67/\sqrt{500} = \hat{p}_i \pm 0.075$ have joint coverage probability at least 0.95 regardless of the number of cells. The inflated width is consistent with expectations and seems a reasonable price to pay for robustness against the usual normality assumptions.

$1-\alpha$	n	t	*asymptotic normal*
0.90	50	1.53	$\hat{p}_i \pm 1.00/\sqrt{n}$
	100	1.48	
	200	1.44	
	500	1.41	
	1000	1.40	
	∞	1.36	
0.95	50	1.67	$\hat{p}_i \pm 1.13/\sqrt{n}$
	100	1.62	
	200	1.58	
	500	1.54	
	1000	1.53	
	∞	1.48	
0.95	50	1.99	$\hat{p}_i \pm 1.40/\sqrt{n}$
	100	1.92	
	200	1.87	
	500	1.82	
	1000	1.79	
	∞	1.73	

5 Polynomial Algorithms for Finding Minimal Connected Enclosures on Embedded planar Graphs

The dissertation of Stutzman, co-advised by myself and Donald Ratliff, studied problems of finding minimal connected enclosures of points or regions in connected, non-separable planar graphs $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and whose edges have non-negative lengths and regions have non-negative weights. The following three problems were considered: (*a*) *Find a shortest enclosing walk for a set of points*. Our algorithm solved the problem in $O(|\mathcal{V}| \log |\mathcal{V}|)$ time by finding a minimum cut on the *dual* graph that has a vertex inside each region of $\mathcal{G}$ and an edge for every pair of adjacent regions. Thus it improved substantially over $O(|\mathcal{V}|^2 \log |\mathcal{V}|)$ algorithms in Bienstock and Monma [6] and Provan [11]. (*b*) *Find a shortest enclosing cycle for a set of points*. This problem differs from the problem in (*a*) in that a cycle cannot have repeated vertices. Stutzman noticed that a polynomial algorithm in Provan [11] does not guarantee an optimal solution. My joint work with Provan and Stutzman has resulted in a correct polynomial algorithm. (*c*) *Find a set of connected regions with minimum total weight that encloses a set of points*. We proved that the problem is *NP–hard* and proposed a polynomial heuristic based on solving a minimum cost flow problem on an "auxiliary" network. Our results will be contained in a paper that is in its final developmental stage.

6 Ongoing Research

The main focus of my current research remains the analysis and design of stochastic networks. Bruce Shultes, a doctoral student whose research is funded by this grant, is studying three different topics:

Estimation of Performability Measures via Markov Chains Traditional Monte Carlo methods are based on drawing independent observations from a distribution on the network state space and evaluating a function $\phi(x)$ for each sample x. Markov chains induce correlation between observations but also offer potential savings in evaluating $\phi(x')$ when the chain makes a transition from x to x'. Recent advances on convergence rates of ergodic Markov chains to their stationary distribution will be will play a major role in our research.

Variance Reduction Methods for Simulating Highly Dependable Systems with Repairs Simulation methods for systems with highly reliable components have been a major research topic for the communications industry (IBM, ATT, etc.) during the last 6 years. We intend to study *importance sampling* methods that take advantage

of structural properties of these systems. Such methods have been very successful for simulating non-repairable systems. I anticipate that the Air Force laboratories and the airline industry will be interested in our results.

Incorporation of Reliability Constraints in Network Design Problems The ultimate goal of resarch in the area of network reliability is to offer design engineers procedures that enhance their ability to design systems for which reliability is an important consideration. Ideally, one would like to generate design models and algorithms which take as input component characteristics and design requirements and produce an "optimal" network design. Since explicit performability expressions for a network are very complex, typical design models replace explicit performability expressions by "surrogates." Recent work has looked at the issue of incorporating probabilistic connectivity constraints into network design models. We intend to study problems related to the incorporation of more general performability constraints.

References

[1] C. Alexopoulos. Distribution-free confidence intervals for conditional probabilities and ratios of expectations. *Management Science* (to appear).

[2] C. Alexopoulos. A note on state space decomposition methods for computing performance measures of stochastic flow networks. *IEEE Transactions on Reliability* (to appear).

[3] C. Alexopoulos. State space partitioning methods for stochastic shortest path problems, Technical Report, School of Industrial and Systems Engineering, Georgia Institute of Technology (submitted for publication).

[4] C. Alexopoulos, A A. El-Tannir, and R. F. Serfozo. A multivariate generalization of Markov modulated processes, Technical Report, School of Industrial and Systems Engineering, Georgia Institute of Technology (submitted for publication).

[5] C. Alexopoulos and A. W. Seila. Conservative confidence intervals for multinomial probabilities. *Operations Research Letters* (to appear).

[6] D. Bienstock and C. L. Monma. Optimal enclosing regions in planar graphs. *Networks*. 19:79–94, 1989.

[7] P. Doulliez and E. Jamoulle. Transportation networks with random arc capacities. *R.A.I.R.O.*, 3:45–60, 1972.

[8] A. A. El-Tannir. *Markov Interactive Processes*. PhD Thesis, School of Industrial and Systems Engineering, Georgia Institute of Technology, 1992.

[9] S. Fitzpatrick and A. Scott. Quick simultaneous confidence intervals for multinomial proportions. *Journal of the American Statistical Association*, 82:875–878, 1987.

[10] J. A. Jacobson. *State Space Partitioning Methods for Solving a Class of Stochastic Network Problems*. PhD Thesis, School of Industrial and Systems Engineering, Georgia Institute of Technology, 1993.

[11] J. S. Provan. Shortest enclosing walks and cycles in embedded graphs. *Information Processing Letters*, 30:119–125, 1989.

[12] B. R. Stutzman. *Zone Formation Problems on Embedded Planar Graphs*. PhD Thesis, School of Industrial and Systems Engineering, Georgia Institute of Technology, 1992.

STATE SPACE PARTITIONING METHODS FOR STOCHASTIC SHORTEST PATH PROBLEMS

CHRISTOS ALEXOPOULOS

Technical Report J-93-01

April 1993

School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332-0205

This research was supported by the Air Force Office of Scientific Research under grant F49620-93-1-0043. Reproduction in whole or part is permitted for any purpose of the United States Government.

Abstract

This paper describes methods for computing measures related to shortest paths in networks with discrete random arc lengths. These measures include the probability that there exists a path with length not exceeding a specified value and the probability that a given path is shortest. The proposed methods are based on an iterative partition of the network state space and provide bounds that improve after each iteration and eventually become equal to the respective measure. These bounds can also be used for constructing simple variance reducing Monte Carlo sampling plans, making the proposed algorithms useful for large problems where exact evaluations are virtually impossible. The algorithms can be easily modified to compute performance characteristics in stochastic activity networks. Computational experience has been encouraging as we have been able to solve networks that have presented difficulties to existing methods.

Key words: Shortest path, reliability, stochastic networks, Monte Carlo methods.

1. Introduction

This paper describes exact and approximation methods for computing measures related to shortest paths in probabilistic networks. Let $G = (N, A, s, t)$ be a directed network with node set $N = \{1,..., t\}$, arc set $A = \{1,..., a\}$, source $s = 1$ and terminal node t. If each arc has a deterministic length (or time required for its traversal), then a basic problem in network optimization is the determination of an (s, t) path with minimum length. Another problem is the identification of the arcs that belong to shortest paths. These problems can be easily solved by a variety of algorithms (see Gallo and Pallottino 1988).

Unfortunately, the presence of uncertainties (such as link failures, variable travel times due to congestion, etc.) in a variety of distribution systems makes a stochastic network model more realistic. These models have applications in the study of communication systems, emergency service delivery systems, spread of fire in a building (see Ling and Williamson 1982), etc. Specifically, we assume that each arc $i \in A$ has a discrete positive random length X_i taking values $x_i(1) < x_i(2) < \cdots < x_i(n_i)$ with respective probabilities $p_i(1), p_i(2),..., p_i(n_i)$. The state space Ω of the random vector $X = (X_1,..., X_a)$ consists of the a-tuples $x(\omega) = (x_1(\omega_1),..., x_a(\omega_a))$, where the index $\omega_i \in \{1,..., n_i\}$ for $i \in A$. To simplify the notation, we use the index k to designate the value $x_i(k)$ for each arc i and level k. Therefore, the state $x(\omega)$ will be denoted by $\omega = (\omega_1,..., \omega_a)$. We consider subsets of Ω that are (discrete) multi-dimentional intervals in the sense that each such set R has lower and upper limiting states $\alpha = \alpha[R]$ and $\beta = \beta[R]$ and all states $\omega \in R$ satisfy $\alpha_i \le \omega_i \le \beta_i$ for all $i \in A$. We denote R by $(\alpha; \beta)$.

Let $L_P(\omega)$ be the length of an (s, t) path P. Also, let $T(\omega)$ denote an s-rooted shortest path tree for the state ω. The predecessor of node j on the tree is denoted by $\ell(j; \omega)$ and the shortest (s, j) path length is denoted by $d(j; \omega)$. Then the path (s, t) $P(\omega)$ with arcs in $T(\omega)$ has length $d(t; \omega)$ and the shortest path length is the random variable

$$L = L(X) = d(t; X). \tag{1}$$

Let $f(\cdot)$ and $F(\cdot)$ denote the probability and cumulative distribution functions of L. Hereafter the term *path* will denote an (s, t) path unless stated otherwise.

For convenience, we assume that the arc lengths are independent random variables. This assumption does not ease the computation of the distribution of $L = \min_{P \in \mathcal{P}} L_P(X)$ by path enumeration because the cardinality of the set of paths $\mathcal{P}$ can be as large as $\lfloor (n-2)!e \rfloor$ for a complete graph on n nodes and the path lengths $L_P(X)$ are not generally independent due to shared arcs. Note that the probability of an interval $R = (\alpha; \beta)$ can be written as

$$P(R) = \prod_{i=1}^{a} \sum_{k=\alpha_i}^{\beta_i} p_i(k). \tag{2}$$

We focus on the computation of the following:

- $F(r)$ = probability that there exists an (s, t) path of length $\leq r$ for fixed r. (3)
- The distribution function $F(\cdot)$ and, therefore, the moments of L.
- $c(e)$ = probability that arc e is in a shortest path. (4)
- $c(P_0)$ = probability that the path P_0 is shortest. (5)

The probabilities $c(e)$ and $c(P_0)$ are often called the *criticality indices* of the arc i and the path P_0. Note that $c(P_0) \leq c(i)$ for all $i \in P_0$. The evaluations of (3)-(5) are $\#P$-hard problems. Indeed, the directed (s, t) reliability problem (see Ball 1987) is the evaluation of $F(0)$ when the arcs assume lengths 0 or 1 with respective probabilities p_i and $1 - p_i$. The following proposition establishes the complexity of the last two evaluations.

Proposition 1: The evaluations of $c(e)$ and $c(P_0)$ for fixed arc e and path P_0 are $\#P$-hard problems.

Proof: Add an arc $e = (s, t)$ with fixed length r. Obviously, e belongs to a shortest path if and only if the shortest (s, t) path in the original network has length $L \geq r$. Since the evaluation of $P(L \geq r)$ is a $\#P$-complete problem, the evaluation of $c(e)$ is a $\#P$-hard problem. Further, for the path $P_0 = \{e\}$ we have $c(P_0) = c(e)$. □

A related problem of considerable interest is the evaluation of the distribution of the length of a longest path in (acyclic) PERT networks. For our setting, Hagstrom (1988, 1990) showed that this problem is $\#P$-complete and proposed a method for solving it.

One possible way for computing the distribution of L is by formulating the problem as a stochastic linear program with random objective coefficients. Bereanu (1966 a,b) and Eubank *et al.* (1974) proposed methods for computing the distribution of the optimal objective value when the coefficients are continuous random variables. These method require the evaluation of the probability that a given basis is optimal; a task that requires a complicated partition of the state space of the objective function. Frank (1969) and Sigal *et al.* (1980) presented exact methods both of which rely on the evaluation of multiple integrals. Because of the great complications that arise in those evaluations, they suggest Monte Carlo sampling. Kulkarni (1986) considered the case of independent and exponentially distributed arc lengths and proposed a method for computing the distribution of L that is based on a Markov process with an absorbing state.

Several studies have addressed the case of independent discrete arc lengths. The approach in Mirchandani (1976) starts with sorting all (s, t) paths and creates a disjoint Boolean expression by comparing neighboring paths. This expression is then used for computing the probability $F(r)$ or the mean of the shortest path length. Hagstrom and Kumar (1984) considered the model where each arc has two modes, an operating mode with

known finite length and a failed mode, and proposed an algorithm for computing the probability $F(r)$. At each step the algorithm conditions (*pivots*) on the mode of a carefully chosen arc to partition the current problem into two simpler subproblems. In the case of multi-state arc lengths they recommended the reduction in Mirchandani (1976) which replaces every arc i with $n_i > 2$ by n_i arcs $i_1,\ldots, i_{n_i}$ where arc i_j is operative with length $x_i(j)$ and probability $p_i(j)/[1 - \Sigma_{k=1}^{j-1} p_i(k)]$ or failed. Unfortunately, this popular reduction increases the numbers of arcs and states in the network.

Corea and Kulkarni (1988a,b) proposed a methodology for computing the distribution of L and criticality indices of paths. They assumed that the arc lengths are integer-valued, replaced each arc with largest possible length m by a subnetwork with $2m$ arcs, and constructed Markov chains with absorbing states and binary transition costs. The above measures are computed by evaluating the distribution of the total cost incurred until absorption. Their construction limits the applicability of the methods to problems of small size.

Hayhurst and Shier (1991) proposed a method for computing the distribution $F(\cdot)$ that is based on *structural factoring*. At each stage the method removes a node and replaces its incident arcs by new arcs. This replacement requires conditioning on the lengths of several incident arcs, identified as factoring arcs. The approach has been successful previously analyzed only by approximation techniques and will serve for comparisons with our methods.

The intractability of the problems under consideration has motivated the development of approximation methods. Several methods, particularly for the case of acyclic networks, rely on bounds as Dodin (1985), Fulkerson (1962), Kleindorfer (1971), and Shogan (1977). Other methods as Sigal *et al.* (1979) and Fishman (1985) rely on Monte Carlo sampling.

Our methods perform an iterative partition of the network state space Ω. They

differ from partition methods designed for the computation of performance measures for flow networks with random capacities (see Doulliez and Jamoulle 1972, Rueger 1986, and Shogan 1982) in their domain of application and approach for partitioning subsets of Ω. Typical performance measures of stochastic networks are the probabilities of certain events related to structural constraints. Examples for our model are the events $\{L(X) \leq r\}$ with probability $F(r)$, $\{L_{P_0}(X) = L(X)\}$ with probability $c(P_0)$, and $\cup_{P:e\in P}\{L_P(X) = L(X)\}$ with probability $c(e)$.

In short, the probability of an event E is computed as follows: We start with the interval Ω. At each iteration an interval $R = (\alpha; \beta)$ is partitioned by solving the knapsack problem

$$\text{maximize} \qquad |\{i \in S: \omega_i = \beta_i\}| \tag{6a}$$

$$\text{subject to} \qquad \sum_{i\in S} x_i(\omega_i) \leq b \tag{6b}$$

$$\omega_i \in \{\alpha_i, \alpha_i + 1, \ldots, \beta_i\} \qquad i \in S, \tag{6c}$$

where S is a subset of a path and b is a positive bound. We solve this problem by renumbering the arcs in S such that $S = \{1', \ldots, q'\}$ with $x_{j'}(\beta_{j'}) - x_{j'}(\alpha_{j'}) \leq x_{k'}(\beta_{k'}) - x_{k'}(\alpha_{k'})$ for each $1 \leq j < k \leq q$ (we break ties by giving preference to larger $\beta_{j'} - \alpha_{j'}$). Then, as long as the constraint (6b) remains valid, for $j = 1, \ldots, q$ we increase the level of arc j' as much as possible.

An optimal solution $\{\gamma_i, i \in S\}$ to problem (6) defines the interval

$$D = \{\omega \in R : \alpha_i \leq \omega_i \leq \gamma_i \text{ for } i \in S\} \tag{7}$$

which either is a subset of E or does not intersect E. In the former case D is called *feasible* and $P(D)$ is part of $P(E)$ while in the latter case E is called *infeasible*. The

remainder $R - D$ is the union of the sets $B'_i = \{\omega \in R: \omega_i > \gamma_i\}$, $i \in S$. These sets are called *undetermined* because their states cannot be classified without computing a new shortest path tree. Since the sets B'_i overlap, we use the well-known approach to partition $R - D$ into the intervals $B_i = B'_i - \cup_{k<i} (B'_i \cap B_k)$ or

$$
\begin{aligned}
B_i = \{\omega \in R:\ & \alpha_k \le \omega_k \le \gamma_k && \text{for } k < i,\ k \in S \\
& \alpha_k \le \omega_k \le \beta_k && \text{for } k < i,\ k \notin S \\
& \gamma_i + 1 \le \omega_i \le \beta_i \\
& \alpha_k \le \omega_k \le \beta_k && \text{for } k > i\} \qquad i \in S \text{ with } \gamma_i < \beta_i .
\end{aligned} \tag{8}
$$

Clearly, an optimal solution to problem (6) minimizes the number of the resulting undetermined intervals. It is also a heuristic solution to the problem of maximizing the number of states in D. These intervals are partitioned in subsequent iterations and the procedure terminates when no undetermined sets remain.

The undetermined intervals are then maintained in a list, say $\mathscr{L}$, whose records consist of the boundary states and, possibly, some additional information. Note that the partition of an interval can be thought of as factoring conditional on the states of the arcs in the corresponding set S.

Remark 1: The set $R - D$ was partitioned into the intervals B_i by considering the overlapping sets B'_i in ascending order of their indices in S. An alternative partition is obtained by considering B'_i in descending order of $i \in S$. Since the determination of an ordering that results in improved long-term performance seems to be a hard problem, we adopt the ordering in (8).

The main advantages of the proposed methods are: (1) Their effectiveness for problems that have presented difficulties to existing methods. (2) Their ability to provide bounds that improve after each iteration and can be used for designing Monte Carlo sampling plans. These plans are conceptually simple, provide estimates with considerably

smaller variances, and require less time per replication than crude Monte Carlo sampling plans. This property makes the methods beneficial for large size problems. (3) Their flexibility for performing sensitivity analysis on the measures of interest with respect to alternative arc length distributions with common state space. (4) Their potential for accommodating statistical dependencies between arc lengths; see the method in Le and Li (1989) for flow networks with dependent arc capacities. (5) Their applicability to problems in stochastic PERT networks after a few modifications.

Section 2 describes the computation of $F(r)$ at fixed r. Section 3 discusses the computation of the distribution function $F(\cdot)$ and proposes a separate algorithm for evaluating the mean $E(L)$ only. Section 4 describes methods for computing criticality indices and section 5 applies our ideas to stochastic activity networks. Each of these sections describes exact algorithm(s), shows how the algorithms produce bounds, and gives Monte Carlo estimators that are based on a variance reducing sampling scheme. Section 6 contains results and section 7 contains final remarks and conclusions.

2. Computing the Probability $F(r)$ for Fixed r

There are numerous ways for decomposing an interval. Clearly, there is a trade-off between obtaining a large feasible subset and the required computational effort. We propose an approach that is conceptually simple and requires at most one shortest path evaluation per iteration. We then describe the overall algorithm and strategies for keeping the number of undetermined intervals low and producing tight bounds quickly.

An undetermined interval $R = (\alpha; \beta)$ is partitioned as follows: We start with a shortest path tree for the lower boundary state α with predecessor labels $\ell(j)$. These labels are used for computing the shortest lengths $d(j)$ and a shortest path P in $O(|N|)$ time. If $d(t) > r$, then all states in R are infeasible. If $d(t) \leq r$, then any state $\omega \in R$ with $\Sigma_{i \in P} x_i(\omega_i) \leq r$ satisfies $L(\omega) \leq r$. The interval D in (7) obtained by solving problem (6) with $S = P$ and $b = r$ is therefore feasible while the states of the

undetermined intervals B_i in (8) cannot be classified without determining a new shortest path.

Note that the lower boundary state of B_i is equal to that of R except for the ith coordinate of B_i which equals $\gamma_i - \alpha_i + 1$. The original shortest path tree can then be used for computing new node labels $\bar{\ell}(j)$ and distances $\bar{d}(j)$ for the state $\alpha[B_i]$. Clearly, B_i is infeasible if $\bar{d}(t) > r$. If $\bar{d}(t) \leq r$, the labels $\bar{\ell}(j)$ are stored along with the boundary states of B_i for use during the partition of this set. This approach increases the storage requirements but results in computational savings. It should be mentioned here that decomposition methods for flow problems compute an entirely new flow in each iteration.

Algorithm PARTITION describes the evaluation of $F(r)$. As mentioned previously, each record in the list $\mathcal{L}$ contains the boundary states of an undetermined set and the optimal node labels corresponding to the lower boundary point. The maintenance of this list is discussed following the algorithm. $F_\ell(r)$ and $F_u(r)$ are lower and upper bounds on $F(r)$ and are updated after the partition of an interval.

ALGORITHM PARTITION(r)

1. Start with the interval $R = \Omega$, empty list $\mathcal{L}$, $F_\ell(r) = 0$, and $F_u(r) = 1$. Compute an s-rooted shortest path tree with arc lengths $x_i(1)$. Let $\ell(j)$ be the predecessor of node j on the tree and let $d(j)$ be the shortest (s, j) path length. If $d(t) > r$, terminate with $F(r) = 0$.
2. Identify a shortest path P.
3. a. Solve problem (6) and compute the feasible interval D in (7) and the undetermined intervals B_i in (8).
   b. Set $F_\ell(r) = F_\ell(r) + P(F)$.
   c. For $i \in P$ with $\gamma_i < \beta_i$:
      - Increase the length of arc i to $x_i(\gamma_i + 1)$ and compute a shortest path tree

with labels $\bar{\ell}(j)$ and shortest (s, j) path lengths $\bar{d}(j)$.

- If $\bar{d}(t) \leq r$, file the record $\{\alpha[B_i]; \beta[B_i]; (\bar{\ell}(j), j \in N)\}$ in $\mathscr{L}$.
- If $\bar{d}(t) > r$, the set B_i is infeasible; set $F_u(r) = F_u(r) - P(B_i)$.

4. If $\mathscr{L}$ is not empty, remove a record $\{\alpha; \beta; (\ell(j), j \in N)\}$ from it and go to step 2.
5. End with $F_\ell(r) = F_u(r) = F(r)$.

Remark 2: The maintenance of $\mathscr{L}$ is an important issue. If the algorithm is to be carried to completion, then $\mathscr{L}$ is maintained as a singly-linked list where depth-first search is carried out to keep the number of records stored low at any time. If on the other hand the objective is the computation of bounds, this list is maintained by using a heap whose nodes are weighed by the probabilities of the intervals and the root with the largest weight is removed in step 4. We use these strategies for the computation of the measures in sections 3-5.

Remark 3: To minimize the number of undetermined intervals, we attempt to force as many arcs with fixed length within the present interval in a shortest path as possible by decreasing their lengths by an $\epsilon > 0$ chosen so that the resulting shortest path remains shortest when these lengths are reset to their original values.

2.1 Monte Carlo Sampling

The basic Monte Carlo method involves sampling from the state space Ω with probabilities $p_i(k)$ for arc i. Suppose that we draw n independent samples $X^{(1)},\ldots, X^{(n)}$. Then

$$\bar{F}(r) = \frac{1}{n} \sum_{j=1}^{n} 1(L(X^{(j)}) \leq r), \tag{9}$$

where $1(\cdot)$ is the indicator function, is an unbiased estimator of $F(r)$ with variance

$$\text{var } \hat{F}(r) = F(r)[1 - F(r)]/n\,. \tag{10}$$

Algorithm PARTITION provides us with the capability of designing a Monte Carlo sampling plan that combines importance and stratified sampling. Indeed, suppose that we decide to exit when the list $\mathscr{L}$ contains k sets, say $U_1,\ldots, U_k$. Let

$$Q_{ij} = \sum_{l=\alpha_i[U_j]}^{\beta_i[U_j]} p_i(l) \qquad i \in A$$

and write

$$\pi_j = P(U_j) = \prod_{i \in A} Q_{ij} \qquad j = 1,\ldots, k\,.$$

We use the *proportional allocation rule* to draw $m_j = n\pi_j/(\Sigma_{j=1}^{k} \pi_j)$, $j = 1,\ldots, k$ independent samples from each U_j as follows:

ALGORITHM SAMPLE($U_1,\ldots, U_k$)

For $j = 1,\ldots, k$:

For $q = 1,\ldots, m_j$:

For $i \in A$: Sample the index ω_i with probabilities $\{p_i(l)/Q_{ij},\ \alpha_i[U_j] \le l \le \beta_i[U_j]\}$ and assign length $X_i^{(j;q)} = x_i(\omega_i)$ to arc i.

Then,

$$\hat{F}(r) = F_{\ell}(r) + \sum_{j=1}^{k} \pi_j(S_j/m_j), \tag{11a}$$

where

$$S_j = \sum_{q=1}^{m_j} 1(L(X^{(j;q)}) \le r), \tag{11b}$$

is also an unbiased estimator of $F(r)$ with variance smaller than that of the crude estimator $\overline{F}(r)$ by a factor of at least

$$1/\left[\sqrt{F_u(r)[1 - F_\ell(r)]} - \sqrt{F_\ell(r)[1 - F_u(r)]}\right]^2.$$

Remark 4: The latter plan requires less mean time per replication than crude Monte Carlo sampling because each set U_j has fewer states than the state space Ω. Also, each U_j must be retained only until all m_j samples are drawn from it and then can be discarded. Therefore these sets can be read one-at-a-time from a file.

A confidence interval for $F(r)$ can be computed by using the central limit theorem or the method in Fishman (1991).

3. Computing the Distribution of L

We discuss two approaches for partitioning an interval. The first applies to the evaluation of the distribution of L, and therefore its moments, and the second to the evaluation of the mean $E(L)$ only.

The first method decomposes an undetermined interval $R = (\alpha; \beta)$ similarly to the method in section 2 with the exception that the indices γ_i for arcs in the shortest path $P(\alpha)$ are set to α_i. All states of the interval

$$D = \{\omega \in R : \omega_i = \alpha_i \text{ for } i \in P\} \tag{12}$$

have $P(\alpha)$ as a shortest path and $P(D)$ is part of the probability $f(d(t; \alpha)) = P[L = d(t; \alpha)]$. We finish by partitioning $R - D$ into the undetermined intervals B_i given in (8) for $\gamma_i = \alpha_i$ and $S = P$. The present shortest path tree $T(\alpha)$ is then used to compute a shortest path tree corresponding to the lower boundary state of each B_i and the record

$\{\alpha[B_i]; \beta[B_i]; (\ell(j), j \in N)\}$ with the updated labels $\ell(j)$ is stored for later consideration.

After each stage the decomposition algorithm produces the lower bounds

$$f_\ell(r) = \Sigma_D P(D), \tag{13}$$

where the sum is over all intervals D whose shortest path length equals r, and the upper bounds

$$f_u(r) = f_\ell(r) + \Sigma_j P(U_j), \tag{14}$$

where U_j are the remaining undetermined intervals.

Algorithm SAMPLE can be used for estimating $f(\cdot)$ when the partitioning procedure terminates with remaining undetermined intervals $U_1,\ldots, U_k$. The probability $f(r)$ is estimated by

$$\hat{f}(r) = f_\ell(r) + \sum_{j=1}^{k} \pi_j(S_j(r)/m_j), \tag{15a}$$

where

$$S_j(r) = \sum_{q=1}^{m_j} 1(L(X^{(j,q)} = r). \tag{15b}$$

A special decomposition of an interval can be performed when we want to compute the mean $E(L)$ only and the probabilities $p_i(k)$ are not decreasing in k for all arcs i. In this case we compute a shortest path P at the most probable state v in R (in case of ties we choose the smallest state) and use the fact that each state of the interval

$$D = \{\omega \in R : \alpha_i \le \omega_i \le v_i \text{ for } i \in P,\ v_i \le \omega_i \le \beta_i \text{ for } i \notin P\} \tag{16}$$

has P as a shortest path. Then the expected length $E(L_P; D)$ of P in the set D is part of $E(L)$ and can be easily computed by

$$E(L_P;\, D) = \sum_{\omega \in D} \left[\sum_{i \in P} x_i(\omega_i) \right] P\{X = x(\omega)\}$$
$$= \prod_{i \notin P} q_i \left[\sum_{i \in P} h_i \cdot \prod_{\substack{j \in D \\ j \neq i}} q_j \right] = P(D) \sum_{i \in P} h_i / q_i , \tag{17a}$$

where

$$q_i = \sum_{\ell = \alpha_i}^{v_i} p_i(\ell) \quad \text{and} \quad h_i = \sum_{\ell = \alpha_i}^{v_i} x_i(\ell) p_i(\ell). \tag{17b}$$

The difference $R - D$ is partitioned into the following undetermined intervals

$$\begin{aligned} C_i = \{\omega \in R:\ & \alpha_k \le \omega_k \le v_k && \text{for } k < i\,,\ k \in P \\ & v_k \le \omega_k \le \beta_k && \text{for } k < i\,,\ k \notin P \\ & \ell_i \le \omega_i \le u_i \\ & \alpha_k \le \omega_k \le \beta_k && \text{for } k > i\}, \end{aligned} \tag{18}$$

where $\ell_i = v_i + 1$, $u_i = \beta_i$ for $i \in P$ with $v_i < \beta_i$ and $\ell_i = \alpha_i$, $u_i = v_i - 1$ for $i \notin P$ with $\alpha_i < v_i$. It should be noted here that the sets C_i and R typically have very distinct most probable states. Since the present shortest path tree $T(v)$ contains little information that might be of use in the computation of a shortest path tree for the most probable state of C_i, the records in the list $\mathcal{L}$ contain only boundary states and a new shortest path is computed at the beginning of an interval partition.

At each stage the partitioning algorithm produces the lower bound

$$E_\ell(L) = \Sigma_D\, E(L_P;\, D), \tag{19}$$

where the sum is taken over all sets D that have been produced. Further, a Monte Carlo sampling plan based on the remaining undetermined sets $U_1, \ldots, U_k$ yields the estimator

$$\hat{E}(L) = E_{\ell}(L) + \sum_{j=1}^{k} \pi_j(S_j/m_j) \tag{20a}$$

with

$$S_j = \sum_{q=1}^{m_j} L(X^{(j;q)}). \tag{20b}$$

4. Computing Criticality Indices

We first describe a method for computing the probability $c(P_0)$ that the path P_0 is shortest. The evaluation of the criticality index of an arc will follow.

The undetermined list $\mathscr{L}$ contains records of the form $\{\alpha; \beta; (\ell(j; \alpha), j \in N)\}$. The partition of the interval $R = (\alpha; \beta)$ proceeds as follows: If P_0 is a shortest path for the state α (equivalently, if $d(j) = d(i) + x_k(\alpha_k)$ for all $k = (i, j) \in P_0$), we compute the length of the second shortest (s, t) path, say $d_2(t)$. Note that P_0 is a shortest path for each state $\omega \in R$ such that $L_{P_0}(\omega) \leq d_2(t)$. We obtain a feasible interval $D = \{\omega \in R : \alpha_i \leq \omega_i \leq \gamma_i \text{ for } i \in P_0\}$ by solving problem (6) with $S = P_0$ and $b = d_2(t)$. The probability of D is added to the current value of $c(P_0)$, the set $R - D$ is decomposed into intervals B_i in (8), and the records $\{\alpha[B_i]; \beta[B_i]; (\bar{\ell}(j), j \in N)\}$ with the updated labels $\bar{\ell}(j)$ for the state $\alpha[B_i]$ are filed in $\mathscr{L}$.

Now suppose that a path P is shorter than P_0. In this case any state $\omega \in R$ satisfying $\Sigma_{i \in P-P_0} x_i(\omega_i) < \Sigma_{i \in P_0-P} x_i(\alpha_i)$ causes $L_P(\omega) < L_{P_0}(\omega)$. Therefore, a solution to problem (6) with $S = P - P_0$ and $b = \Sigma_{i \in P_0-P} x_i(\alpha_i) - \epsilon$ (for appropriate $\epsilon > 0$) yields an infeasible interval $D = \{\omega \in R : \alpha_i \leq \omega_i \leq \gamma_i \text{ for } i \in P - P_0\}$. The remainder $R - D$ is partitioned and the new records are filed in $\mathscr{L}$.

After each iteration the overall algorithm produces the bounds

$$c_{\ell}(P_0) = \Sigma_F P(F) \tag{21}$$

and

$$c_u(P_0) = 1 - \Sigma_I P(I), \tag{22}$$

where the first (second) sum is over all feasible (infeasible) intervals that have been obtained. Also, Monte Carlo sampling based on the undetermined sets $U_1,\ldots, U_k$ produces the estimator

$$\hat{c}(P_0) = c_\ell(P_0) + \sum_{j=1}^{k} \pi_j(S_j/m_j), \tag{23}$$

where S_j increases by one at a trial from U_j only if P_0 is a shortest path.

The evaluation of the criticality index of an arc e proceeds similarly to the evaluation of the criticality index of a path. The partition of the interval $R = (\alpha; \beta)$ starts with the identification of a shortest path P corresponding to the state α. To force the arc e in P, the shortest path tree is computed with the length $x_e(\alpha_e)$ reduced by ϵ. If $e \in P$, we solve problem (6) with $S = P$ and $b = d_2(t)$. Each state in the interval $D = \{\omega \in R : \alpha_i \le \omega_i \le \gamma_i \text{ for } i \in P\}$ has P as a shortest path and then $P(D)$ is part of $c(e)$. If $e \notin P$, we solve this problem with $b = d_2(t) - \epsilon$. In this case no state in the resulting interval D has e in a shortest path making the interval infeasible. The partition of R ends with the decomposition of the remainder $R - D$ into the intervals B_i in (8) and the records $\{\alpha[B_i]; \beta[B_i]; (\ell(j), j \in N)\}$ with the updated labels $\ell(j)$ are filed in $\mathcal{L}$.

Finally, the estimation of $c(e)$ by Monte Carlo simulation proceeds similarly to the estimation of $c(P_0)$.

Remark 5: If the network is acyclic, we can obtain a *larger* infeasible subset of R when e is not in a shortest path at the cost of an additional shortest path evaluation. In this case we let $e = (i, j)$ and compute a shortest (j, t) path with length $h(j)$. Then $u_e = d(i) + x_e(\alpha_e) + h(j)$ is the length of a shortest (s, t) path containing e, all states $\omega \in R$ with $L_P(\omega) < u_e$ are infeasible, and the infeasible set results by solving problem (6) with $S = P$ and $b = u_e - \epsilon$.

Remark 6: The evaluations of criticality indices are clearly more time-consuming

than the evaluation of $F(r)$ at a single r. The examples in section 6 will show that the former evaluations have time requirements of the same order of magnitude as the computation of the distribution $F(\cdot)$. Lower bounds for criticality indices can also be obtained during the computation of $F(\cdot)$ by noting that the probability of the interval D in (12) is part of both $c(P)$ and $c(i)$ for $i \in P$.

5. Applications to Stochastic Activity Networks

Appropriate modifications make the methods in sections 2-4 applicable to problems in activity networks with discrete random task durations. These acyclic networks represent projects with tasks corresponding to arcs. All tasks directed into a node must be completed before any task directed out of it can be started. The project is complete when all tasks directed into the terminal node t are finished and the duration of the project is the length of the longest (s, t) path. The nodes of an activity network can be labelled so that $i < j$ for each arc (i, j). For fixed durations $\ell_1, \ldots, \ell_a$ the project duration $d(t)$ can be computed by the following recursion in time $O(|A|)$:

Set $d(s) = 0$.

For $j = 2, \ldots, t$: Set $d(j) = \max_{(i,\, j) \in A} \{d(i) + \ell_{(i,j)}\}$.

We briefly discuss the evaluation of $P(L \geq r)$ for fixed r, where now L denotes the project duration. To partition an undetermined interval $R = (\alpha; \beta)$ we start with its upper boundary state and compute a longest path P. If $L_P(\beta) < r$, the interval R is infeasible; otherwise, we find an optimal solution $\{\delta_i, i \in P\}$ to the problem

$$\begin{aligned} &\text{maximize} && |\{i \in P\colon \delta_i = \alpha_i\}| \\ &\text{subject to} && \sum_{i \in P} x_i(\delta_i) \geq r \end{aligned} \tag{24}$$

$$\delta_i \in \{\alpha_i, \alpha_i + 1, \ldots, \beta_i\} \qquad i \in P$$

which is similar to problem (6). The interval

$$D = \{\omega \in R : \delta_i \le \omega_i \le \beta_i \text{ for } i \in P\} \tag{25}$$

is obviously feasible and the difference $R - D$ is partitioned into the intervals

$$\begin{aligned} H_i = \{\omega \in R : \ & \delta_k \le \omega_k \le \beta_k && \text{for } k < i \\ & \alpha_i \le \omega_i \le \delta_i - 1 \\ & \alpha_k \le \omega_k \le \beta_k && \text{for } k > i\} \end{aligned} \qquad i \in P \text{ with } \delta_i > \alpha_i . \tag{26}$$

Since a longest path can be computed in time linear in the number of arcs, the records in the list $\mathscr{L}$ contain only the boundary states of the respective intervals.

6. Examples

Figure 1 shows a network with 10 nodes, 23 arcs, $s = 1$, and $t = 10$. The numbers on each arc give the probability function of its length. For instance, arc (1, 2) has length 7.0, 7.3 or 9.4 with probabilities 0.2, 0.5 or 0.3 respectively. The programs were written in FORTRAN 77 and run on a SUN SPARCSTATION IPC, and the subroutine L2QUE from Gallo and Pallottino (1988) was used for computing shortest paths. Table 1 lists the results for this network. Note that the evaluation of the $P(L \le 13)$ took 0.05 seconds and required 47 set partitions. In general, all the CPU times in column 2 are less than one second. Also, the computation of the distribution of L required 12330 set partitions and took only 5.57 seconds.

The method in Hayhurst and Shier (1991) computed this distribution in time that was larger by several orders of magnitude (their algorithm was written in PASCAL and

run on a different computer). This method is based on factoring and performs convolutions between arc lengths which are often time-consuming. The method in Corea and Kulkarni (1988a) is practically non-applicable to this problem because it requires integer-valued arc lengths and replaces every arc with maximum possible length m by a subnetwork with $m + 1$ nodes and $2m$ arcs. For example, arc $(1, 2)$ would be replaced by $94 + 1 = 95$ nodes and $2\times94= 188$ arcs. Then the approach constructs a Markov chain with absorbing states. These replacements result in Markov chains with an astronomical number of states. However, it should be noted that the methods of Corea and Kulkarni (1988a,b) appear to be as efficient as our method when the arc lengths are i.i.d. discrete uniform random variables with range $\{1, 2,..., k\}$ for $k \leq 3$. A clear advantage of our approaches is their ability to produce bounds.

Table 1 also lists the criticality indices of selected paths and arcs. Since the network is acyclic, a path is denoted by its sequence of nodes. The paths and arcs that are not listed have small criticality indices.

Table 2 displays results for the same network but with four-state arc lengths. The probability functions are not drastically different from those in Figure 1. Note that the CPU times for computing $P(L \leq r)$ for small r remain significantly shorter than the times required for r in the middle of the range of L. The small size of the lists of undetermined rectangles is attributable to the effectiveness of the LIFO maintenance.

We now consider the network in Figure 2 with 15 nodes and 42 arcs. The number of possible lengths for each arc was chosen at random from the set $\{1, 2, 3\}$, the lengths were generated randomly from $\{10, 11,..., 40\}$, and their respective probabilities were selected from $\{0.1, 0.2,..., 0.9\}$ and were ranked so that the smaller length has the largest probability. The results are listed in Table 3. For instance, the computation of the distribution of L took 228.73 seconds and required the partition of 298918 sets. It should be noted that the latter number is a small fraction (5.91×10^{-15}) of the total number of states. For this problem the time requirements for computing $F(r)$ for fixed r

increase significantly for $r \geq 75$ but the number of sets remains very small relative to the cardinality of the sample space. On the other hand, 1000 partitions with the undetermined sets maintained with the use of heaps produced reasonably tight bounds.

The table also contains criticality indices. As with the network in Figure 1, the evaluation of a path criticality index requires roughly three times as few partitions as the evaluation of an arc criticality index.

The results in Table 4 illustrate the method in section 3 for estimating the distribution of L. The bounds in columns 2 and 3 resulted after only 10000 intervals were partitioned with the undetermined intervals processed via a heap with root corresponding to the most probable set. Note that the computation of $F(r)$ for $r \leq 54$ was completed before the algorithm was terminated. A total of 20000 samples were drawn from the remaining intervals by using algorithm SAMPLE in section 2. The accuracy of the estimates in column 3 and the large variance reduction ratios versus crude Monte Carlo indicate the overall contribution of the proposed methods.

We finally studied the effectiveness of the strategy in remark 6 in an acyclic network resulting from the network in Figure 2 after a few arcs are reversed. The reduction in the number of partitioned sets by an average of 10-15 percent did not justify the 50 percent increase in the CPU times caused by the additional shortest path evaluation.

7. Conclusions

We proposed a set of methods for computing measures related to shortest paths in networks with random arc lengths. These methods are based on an iterative partition of the system state space and gain their effectiveness from their ability to generate bounds and information that can be used for constructing simple and efficient Monte Carlo sampling plans. These methods can also be applied to the computation of other system characteristics as the conditional probability that the shortest path length exceeds a given value given that a specified arc belongs to a shortest path.

References

Alexopoulos, C. (1989). Distribution-free confidence intervals for conditional probabilities and ratios of expectations, Technical Report series No. J-89-3, School of Industrial and Systems Engineering, Georgia Institute of Technology, revised August 1992.

Ball, M.O. (1987). Computational complexity of network reliability analysis: An overview, *IEEE Transactions on Reliability*, **35**, 230-239.

Bereanu, B. (1966a). On stochastic linear programming: The Laplace transform of the distribution of the optimum and applications, *J. Math. Anal. Appl.*, **15**, 280-294.

Bereanu, B. (1966b). On stochastic linear programming II: Distribution problems: non-stochastic technology matrix, *Rev. Roumaine Math. Pures Appl., II* , 713-725.

Corea, G.A. and V.G. Kulkarni (1988a). Shortest paths in stochastic networks with discrete arc lengths, forthcoming in *Networks.*

Corea, G.A. and V.G. Kulkarni (1988b). Criticality indices of paths in networks with random arc lengths having discrete distributions, Technical Report No. UNC/OR/TR-88/2, Department of Operations Research, University of North Carolina at Chapel Hill.

Dodin, B.M. (1985). Bounding the project completion time distribution in PERT networks, *Operations Research*, **23**, 862-881.

Doulliez, P and E. Jamoulle (1972). Transportation networks with random arc capacities, *R.A.I.R.O.*, **3**, 45-60.

Eubank, J.B., B.L. Foote, and H.J. Kumin (1974). A method for the solution of the distribution problem of stochastic liner programming, *SIAM J. Appl. Math.*, **26**, 225-238.

Fishman, G.S. (1985). Estimating network characteristics in stochastic activity networks, *Management Science*, **31**, 579-593.

Fishman, G.S. (1991). Confidence intervals for the mean in the bounded case, *Statistics and Probability Letters*, **12**, 223-227.

Fulkerson, D.R. (1962). Expected critical path lengths in PERT networks, *Operations Research*, **10**, 808-817.

Gallo, G. and S. Pallottino (1988). Shortest path algorithms, *Annals of Operations Research*, **13**, 3-79.

Hagstrom, J.N. (1988). Computational complexity of PERT problems, *Networks*, 18, 139-147.

Hagstrom, J.N. (1990). Computing the probability distribution of project duration in a PERT network, *Networks*, **20**, 231-244.

Hagstrom, J.N. and P. Kumar (1984). Reliability computation on a probabilistic network with path length criterion, Technical Report, University of Illinois at Chicago.

Hayhurst, K.J. and D.R. Shier (1991). A factoring approach for the stochastic shortest path problem, *Operations Research Letters*, **10**, 329-334.

Kleindorfer, G.B. (1971). Bounding distributions for a stochastic acyclic network, *Operations Research*, **19**, 1586-1601.

Le, K.V. and V.O.K. Li (1989). Modeling and analysis of systems with miltimode components and dependent failures, *IEEE Transactions on Reliability*, **29**, 68-75.

Nemhauser, G.L. and L.A. Wolsey (1988). *Integer and Combinatorial Optimization*, Wiley, New York.

Ling, W.C.T. and R.B. Williamsom (1982). Using fire tests for quantitative risk analysis, *Fire Risk Assessment*, ASMT SPT 762, G.T. Castino and T.Z. Harmathy (Editors), American Society for Testing and Materials, 38-58.

Mirchandani, P.B. (1976). Shortest distance and reliability of probabilistic networks, *Computers and Operations Research*, **3**, 347-355.

Rueger, W.J. (1986). Reliability analysis of networks with capacity constraints and failures at branches and nodes, *IEEE Transactions on Reliability*, **35**, 523-528.

Shogan, A.W. (1977). Bounding distributions for a stochastic PERT network, *Networks*, **7**, 359-381.

Shogan, A.W. (1982). Modular decomposition and reliability computation in stochastic transportation networks having cutnodes, *Networks*, 12, 255-275.

Sigal, L.E., A.A.B. Pritsker, and J.J. Solberg (1980). The use of cutsets in Monte Carlo analysis of stochastic networks, *Math. Comp. Simulation*, **21**, 376-384.

Sigal, L.E., A.A.B. Pritsker, and J.J. Solberg (1980). The stochastic shortest route problem, *Operations Research*, **23**, 1122-1129.

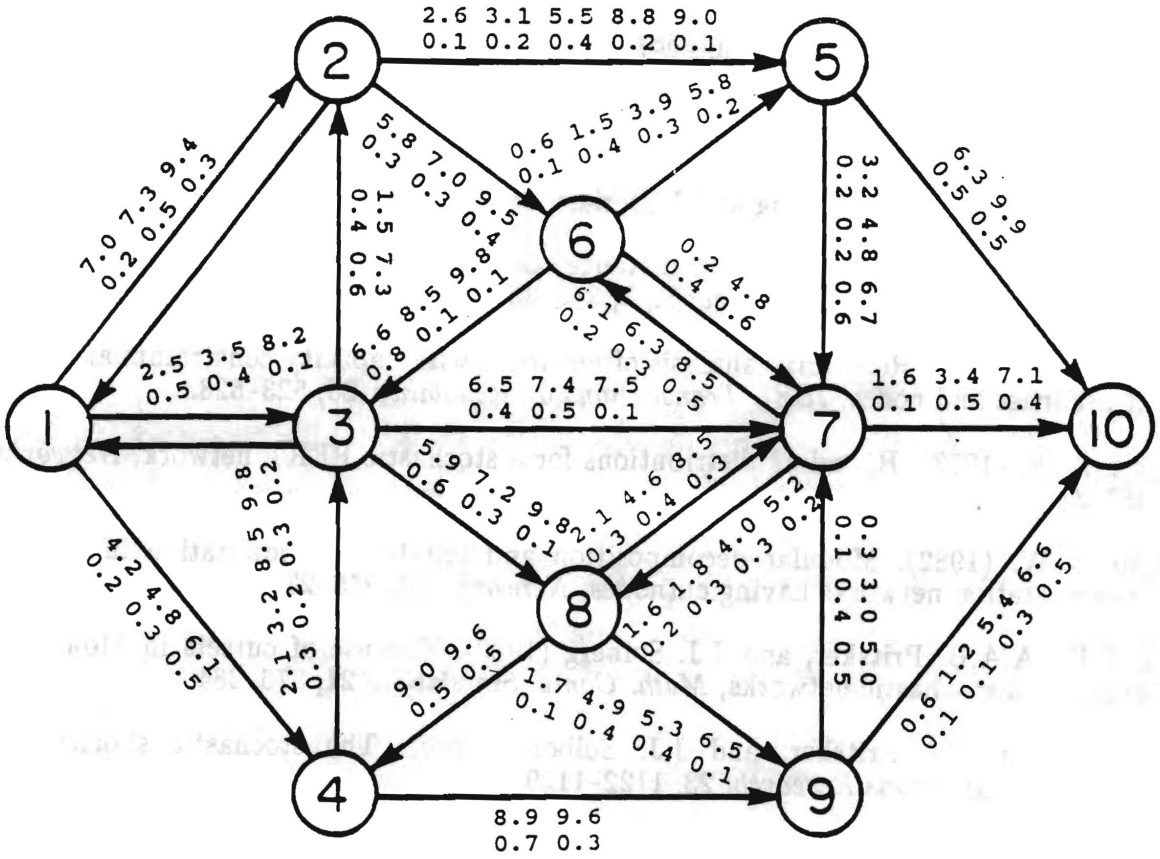
1
2
3
4
5
6
7
8
9
10
2.6 3.1 5.5 8.8 9.0
0.1 0.2 0.4 0.2 0.1
7.0 7.3 9.4
0.2 0.5 0.3
5.8 7.0 9.5
0.3 0.3 0.4
0.6 1.5 3.9 5.8
0.1 0.4 0.3 0.2
3.2 4.8 6.7
0.2 0.2 0.6
6.3 9.9
0.5 0.5
1.5 7.3
0.4 0.6
6.6 8.5 9.8 4.5
0.8 0.1 0.1
0.2 4.8
0.4 0.6
6.1
0.2
6.3 8.5
0.3 0.5
2.5 3.5 8.2
0.5 0.4 0.1
6.5 7.4 7.5
0.4 0.5 0.1
1.6 3.4 7.1
0.1 0.5 0.4
5.9 7.2 9.8
0.6 0.3 0.1
2.1 4.6 8.5
0.3 0.4 0.3
9.8
0.2
8.5
0.3
3.2
0.2
4.2
0.2
4.8
0.3
0.5
7.5
2.1
0.3
4.0 5.2
0.3 0.2
1.8
0.3
1.6
0.2
0.3 3.0 5.0
0.1 0.4 0.5
9.0 9.6
0.5 0.5
1.7 4.9 5.3 6.5
0.1 0.4 0.4 0.1
0.6 1.2 5.4 6.6
0.1 0.1 0.3 0.5
8.9 9.6
0.7 0.3

Figure 1

Table 1

Results for the network in Figure 1

Number of states = 87.07×10^9
Expected shortest path length = 14.70832
Number of partitioned sets = 12330
CPU time = 5.57 seconds

r	CPU seconds	no. of partitioned sets	$P(L \leq r)$
10.6	.03	4	.02000
11.	.04	10	.02288
12.	.05	17	.07505
13.	.05	47	.20859
14.	.13	281	.45722
15.	.21	568	.61813
16.	.22	619	.65679
17.	.20	534	.82858
18.	.13	286	.93966
19.	.21	573	.97381
20.	.15	306	.98498
21.	.09	187	.99348
22.	.09	156	.99922
22.3			1.

At most 11 undetermined sets were stored at any time.

path	CPU seconds	no. of partitioned sets	criticality index
$1 \rightarrow 3 \rightarrow 7 \rightarrow 10$	0.92	2323	.7578
$1 \rightarrow 4 \rightarrow 9 \rightarrow 10$	1.90	4749	.0762

arc	CPU seconds	no. of partitioned sets	criticality index
(1, 3)	2.98	6366	.8686
(1, 4)	3.34	6963	.0977
(2, 5)	3.12	6669	.0968
(3, 2)	3.56	7600	.0655
(3, 7)	3.07	6557	.7727
(4, 9)	3.09	6591	.0837
(7, 10)	2.98	6535	.7869
(9, 10)	3.13	6737	.1316

At most 10 undetermined sets were stored at any time.

Table 2

Results for the network in Figure 1 with four-state arc lengths

Number of states = 70.37×10^{12}
Expected shortest path length = 14.95739
Number of partitioned sets = 66476
CPU time = 30.55 seconds

r	CPU seconds	no. of partitioned sets	$P(L \leq r)$
9.9	.03	4	.01600
11.	.04	25	.04297
12.	.09	149	.11058
13.	.20	551	.22208
14.	.49	1381	.34606
15.	1.02	2930	.54963
16.	2.12	6443	.67617
17.	2.05	6137	.82010
18.	2.40	7108	.90792
19.	1.97	5761	.95920
20.	.82	2240	.98964
21.	.32	866	.99439
22.	.17	407	.99943
22.1			1.

At most 12 undetermined sets were stored at any time.

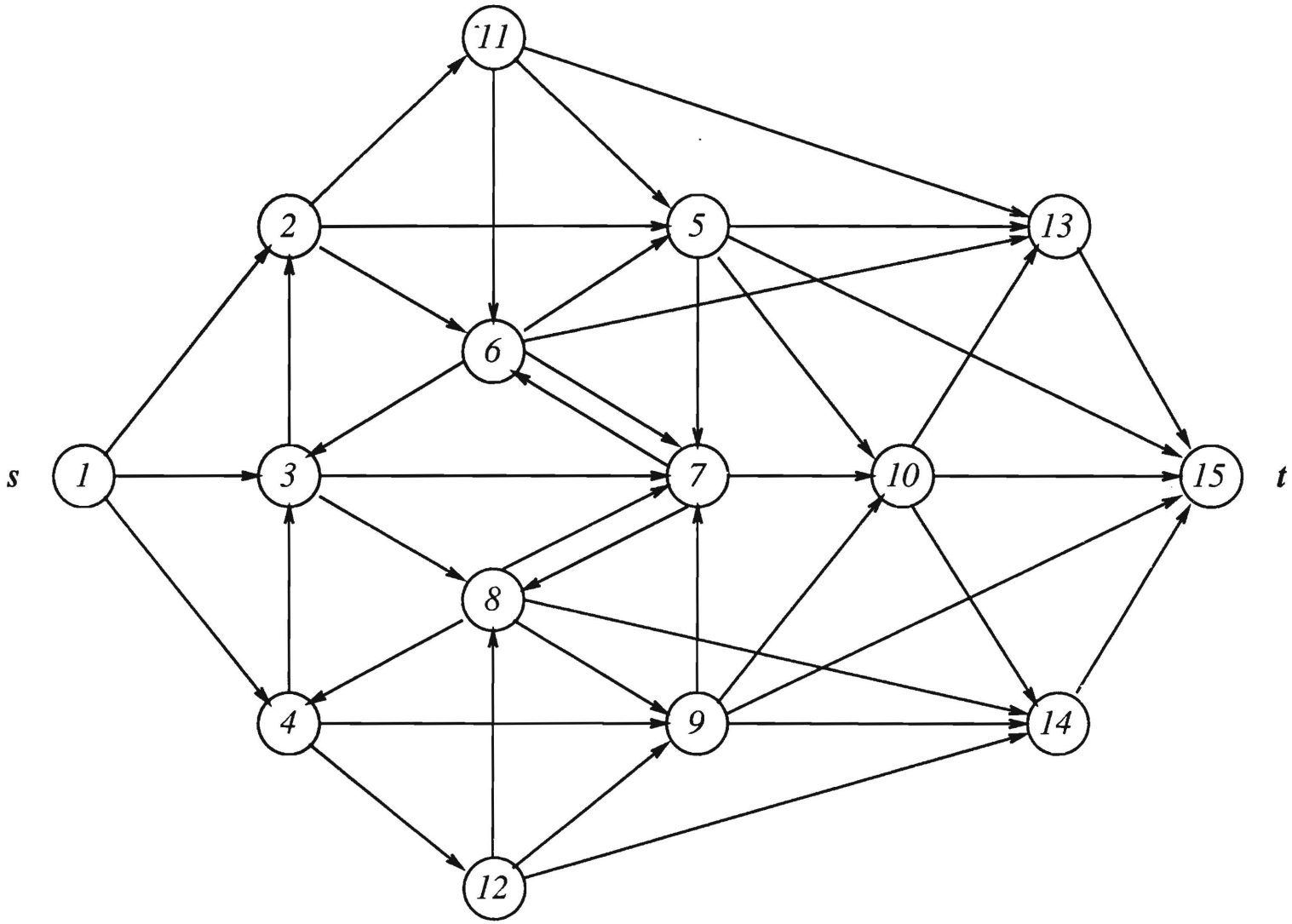

Figure 2

Table 3

Results for the large network in Figure 2

Number of states = 5.05×10^{19}
Expected shortest path length = 57.81583
Number of partitioned sets = 298918
CPU time = 228.73 seconds

r	CPU seconds	no. of partitioned sets	$P(L \leq r)$
51.	.06	5	.03
52.	.07	17	.31518
54.	.07	26	.35754
56.	.09	74	.43363
58.	.11	92	.52863
60.	.15	187	.70651
62.	.33	531	.84400
64.	.48	802	.89552
66.	.35	598	.92236
68.	.61	1130	.94890
70.	.90	1700	.96466
75.	5.12	10072	.99469
80.	6.06	11310	.99940
85.	6.86	13834	.99991
90.	15.52	30345	.99999
95.	8.69	15703	.9999998
98.			1.

At most 12 undetermined sets were stored at any time.

Bounds after 1000 partitions

r	lower bound	upper bound
75.	.89326	.99950
80.	.98803	.99992
85.	.99245	.999997
90.	.99493	.99999997
95.	.99853	.99999996

Table 3 (continued)

Criticality indices for selected paths and arcs

path	CPU seconds	no. of partitioned sets	criticality index
$1 \rightarrow 2 \rightarrow 5 \rightarrow 15$	55.94	89446	.4654
$1 \rightarrow 4 \rightarrow 9 \rightarrow 15$	64.70	103650	.1572
$1 \rightarrow 4 \rightarrow 12 \rightarrow 14 \rightarrow 15$	53.82	85833	.1815
$1 \rightarrow 2 \rightarrow 11 \rightarrow 13 \rightarrow 15$	42.39	67060	.0092

At most 13 undetermined sets were stored at any time.

arc	CPU seconds	no. of partitioned sets	criticality index
(1, 2)	211.70	290500	.5237
(1, 4)	208.26	285712	.4403
(2, 5)	212.56	293052	.4743
(2, 11)	210.57	290075	.0100
(4, 9)	211.00	290735	.1596
(4, 12)	205.43	282893	.2862
(5, 15)	212.13	293052	.4743
(9, 15)	207.84	287035	.3163
(11, 13)	210.14	290075	.0100
(12, 14)	205.96	284245	.1822
(13, 15)	211.49	292200	.0546
(14, 15)	201.22	277850	.1948

At most 14 undetermined sets were stored at any time.

Table 4

Monte Carlo estimation of the distribution of the shortest path length for the network in Figure 2

Sample size $n = 20000$
The bounds were computed after 10000 sets were partitioned by using a heap

r	$F_\ell(r)$	$F_u(r)$	estimate	variance $\times 10^7$	variance† reduction
51.	.03	.03			
52.	.3152	.3152			
54.	.3575	.3575			
56.	.4282	.4534	.4334	.463	321.15
58.	.5105	.5408	.5291	.806	365.52
60.	.6583	.7991	.7066	2.981	50.36
62.	.7388	.9087	.8446	3.775	29.29
64.	.7596	.9661	.8965	3.983	17.45
66.	.7623	.9749	.9226	3.28	16.75
68.	.7747	.9904	.9491	2.623	14.05
70.	.7793	.9957	.9657	2.194	10.41
75.	.7829	.9997	.9948	.496	5.62
80.	.8785	.99997	.9994	.063	5.46
85.	.9782	.999994	.9998	.016	5.16

† The variance reduction ratios were estimated.

A Multivariate Generalization of Markov Modulated Processes

Christos Alexopoulos Akram A. El-Tannir Richard F. Serfozo

December 29, 1993

Abstract

We study a class of multivariate ergodic Markov processes which includes Markov modulated queues and Jackson network processes. We have three objectives. First, we give sufficient conditions for their stationary distribution to have a product form. Secondly, we propose approximations for their stationary distribution. Thirdly, we show how our findings can be used for analyzing the equilibrium behavior of several practical systems.

Authors' addresses: Christos Alexopoulos, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332; Richard F. Serfozo, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332; Akram A. El-Tannir, Arabdar Consultants, P. O. Box 35231 Shaab, 36053 Kuwait.

1 Introduction

Service systems with queueing are often subject to variations in their arrival and service rates. Such variations are due to several factors such as resource allocations, control decisions and unexpected interruptions. During the last three decades, there have been many studies that considered mainly single queues whose arrival and service rates are determined by the state of an extraneous Markov process representing the *environment.* The latter process is not affected by the queueing process. Those models are called "Markov Modulated Processes" (MMP).

The earliest of these is the Markov modulated $M/M/1$ queue in which the environment process alternates between two states and the arrival and service rates are functions of the environment state. This system is described by a two-dimensional Markov process $\{(X(t), Y(t)) : t \geq 0\}$. where $X(t)$ is the number of customers in the system and $Y(t)$ is the state of the environment at time t. This model was first investigated by [4]

[17] and [21]. Later works like [10] [11] [12] [13] [14] [15] [20] studied generalizations of that model to more than two states for the extraneous Markov process, multiple servers, or/and general service times. The name "Markov Modulated Process" appeared first in the early eighties. Since then many studies such as [2] [5] [8] [9] [18] [22] [23] continued to develop these single queueing systems. A survey of the Markov modulated $M/M/1$ and $M/M/s$ models is given in [16].

The main theme of the above studies is to define the process by two components, the *system* component that describes the number of customers, and the environment component that defines the state under which the system operates with the corresponding arrival and service rates. The environment component changes according to a Markov process that is independent of the system component. Those models, however, do not cover the cases where the environment may also be affected by the system as well. For example, the rates of the environment transitions in an $M/M/1$ queue mentioned above may depend on the number of customers in the system at the time of fluctuations. Another example is an $M/M/Y$ system where the number of servers $Y(t)$ at time t is a Markov process with rates that depend on the number of customers in the system.

This paper studies a family of multivariate or (multi-component) Markov processes where transitions can take place simultaneously and the rate at which a set of components changes state depends on the state of the remaining components. To emphasize the interaction between the components, we call these processes "Markov Interactive Processes" (MIP). This family covers a wide range of Markov processes including the MMP's mentioned above, and some standard network processes such as the closed Jackson network process.

We have three primary objectives. Section 2 identifies MIP's whose stationary distribution has a product form. Section 3 presents approximations for stationary distributions without a product form. Section 4 identifies systems that can fit within the framework of MIP's. A few numerical examples demonstrate the efficacy of the proposed approximations.

2 Product Form Distributions for MIP's

We start with the formal definition of an MIP.

Definition 1 *Let* $\boldsymbol{X} = \{(X_1(t), \ldots, X_m(t)) : t \geq 0\}$ *be a continuous time Markov process with a countable state space and let* $\mathcal{K}$ *be a collection of subsets of* $\{1, \ldots, m\}$. *We call* $\boldsymbol{X}$ *a Markov Interactive Process (MIP) if its transition rates are of the form*

$$\lambda(\boldsymbol{x}, \boldsymbol{x}') = \begin{cases} \lambda_K(\boldsymbol{x}_K, \boldsymbol{x}'_K; \boldsymbol{x}^c_K) & \text{for some } K \in \mathcal{K} \text{ and } \boldsymbol{x}'_{K^c} = \boldsymbol{x}_{K^c} \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where $\boldsymbol{x}_K$ *is the vector of* x_j *with* $j \in K$ *and* K^c *is the complement of* K *in* $\{1, \ldots, m\}$.

Note that a transition in $\boldsymbol{X}$ occurs when only the components in a set K change state. The components X_j may be real numbers or take values in a general space. We assume throughout this paper that the MIP $\boldsymbol{X}$ is ergodic and has a stationary distribution denoted by $\pi(\boldsymbol{x})$. The marginal distribution of X_j is denoted by $\pi_j(x_j)$.

For simplicity of exposition, much of our study will focus on a bivariate MIP (X, Y) with rates

$$\lambda(x, y; x', y') = \begin{cases} q(x, x'; y) & x \neq x', y = y' \\ r(y, y'; x) & x = x', y \neq y' \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

We view X as the system component and Y as the environment component. As mentioned in the introduction, MMP's are special cases of bivariate MIP's where the transition rates for component Y are independent of the state of X.

We first consider the case in which the rates $\lambda(\boldsymbol{x}, \boldsymbol{x}')$ have the form

$$\lambda(\boldsymbol{x}, \boldsymbol{x}') = \begin{cases} \frac{\gamma_K(\boldsymbol{x}_{K^c})}{\Phi(\boldsymbol{x})} \prod_{j \in K} q_j(x_j, x_j') & \text{for some } K \in \mathcal{K} \text{ and } \boldsymbol{x}'_{K^c} = \boldsymbol{x}_{K^c} \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where $\gamma_K(\cdot)$ and $\Phi(\cdot)$ are positive functions. For $j = 1, \ldots, m$ define a Markov process $\tilde{X}_j$ with transition rates $q_j(x_j, x_j')$ and state space the same as X_j. Assume that $\tilde{X}_j$ is ergodic and denote its stationary distribution by $\tilde{\pi}_j(x_j)$. The next theorem shows that the stationary distribution of $\boldsymbol{X}$ has a product form.

Theorem 2 *The stationary distribution of the process* $\boldsymbol{X}$ *with transition rates* (3) *is given by*

$$\pi(\boldsymbol{x}) = c\Phi(\boldsymbol{x}) \prod_{j=1}^{m} \tilde{\pi}_j(x_j), \tag{4}$$

where c *is a normalizing constant. If in addition the state space of* $\boldsymbol{X}$ *is the Cartesian product of the state spaces of its components and the function* $\Phi(\cdot)$ *is constant, then the stationary distribution of each* $\tilde{X}_j$ *is equal to the marginal stationary distribution of* X_j *and*

$$\pi(\boldsymbol{x}) = \prod_{j=1}^{m} \pi_j(x_j). \tag{5}$$

Proof Multiplying the balance equations for $\tilde{X}_j$ for $j \in K$ we have

$$\prod_{j \in K} \left[\tilde{\pi}_j(x_j) \sum_{x_j'} q_j(x_j, x_j') \right] = \prod_{j \in K} \left[\sum_{x_j'} \tilde{\pi}_j(x_j') q_j(x_j', x_j) \right].$$

Multiplication of both sides of the latter equations by $c\gamma_K(\boldsymbol{x}_{K^c}) 1(\boldsymbol{x}'_{K^c} = \boldsymbol{x}_{K^c}) \prod_{j \in K^c} \tilde{\pi}_j(x_j)$ shows that $\pi(\cdot)$ in (4) satisfies the partial balance equations for the set

K. Addition of these equations over all $K \in \mathcal{K}$ results in the total balance equations of the process $\boldsymbol{X}$.

When $\Phi(\cdot)$ is constant and the state space of $\boldsymbol{X}$ is a Cartesian product, summation of $\pi(\boldsymbol{x})$ over $\boldsymbol{x}$ shows that the normalizing constant is equal to unity and the product form of $\pi(\boldsymbol{x}) = \prod_{j=1}^{m} \tilde{\pi}_j(x_j)$ implies that $\tilde{\pi}_j(x_j)$ is the stationary distribution of X_j. Equation (5) follows. □

Remark 3 The rates (3) are by no means necessary to obtain a product form stationary distribution for an MIP. A less restrictive case where a product form is obtained is presented in section 4.5.

Example 4 *Closed Markovian Networks with Batch Movements.*
Henderson et al. [6] studied a closed Markovian network with J nodes and N units. The network was modelled by the process $\boldsymbol{X} = \{(X_1(t), X_2(t), \ldots, X_N(t)) : t \geq 0\}$, where $X_j(t)$ is the site (node) of unit j at time t. Units can move simultaneously with transition rates given by (3), where $\gamma_K(\boldsymbol{x}_{K^c})/\Phi(\boldsymbol{x})$ is the probability that the units in the set K are chosen to change sites within K when the system is in state $\boldsymbol{x}$, and $q_j(x_j, x'_j)$ is the probability that unit j will move from site x_j to site x'_j. The stationary distribution of $\boldsymbol{X}$ is then given by (4).

Example 5 *Closed Jackson Network.*
A closed Jackson network with N units can also be modelled by a process $\boldsymbol{X} = (X_1, \ldots, X_N)$ describing the location of the units. A transition occurs when a unit j moves from node x_j to another node x'_j and the transition rates have the form (3), where $\mathcal{K} = \{\{1\}, \ldots, \{m\}\}$, $q_j(x_j, x'_j)$ is the routing intensity for customer j from node x_j to node x'_j, and the ratio $\gamma_j(\boldsymbol{x}_j^c)/\Phi(\boldsymbol{x})$ is the service rate at node x_j. Then $\boldsymbol{X}$ is an MIP and its stationary distribution has the product form (4).

Observe that the intensities $q_j(x_j, x'_j)$ do not depend on x_k for $k \neq j$. Hence, the result in this example also holds for networks with multiple types of units and appropriately defined rates. Finally, modelling open networks in a similar fashion appears to be a difficult task.

Remark 6 A closed Jackson network is usually modelled by the process $\{(Y_1(t), Y_2(t), \ldots, Y_J(t)) : t \geq 0\}$, where Y_j counts the number of units at node j. Note that this process is less informative than $\boldsymbol{X}$ and its stationary distribution can be expressed in terms of the stationary distribution $\pi(\cdot)$ of $\boldsymbol{X}$.

Example 7 *Markov Modulated $M/M/1$ Queue.*

Suppose (X, Y) is an ergodic Markov modulated $M/M/1$ queue that fluctuates between two environments, say 1 and 2. Assume that the transition rates of component X at a fixed environment y are given by

$$q(x, x'; y) = \begin{cases} \lambda\gamma(y) & \text{if } x' = x + 1,\ x \geq 0 \\ \mu\gamma(y) & \text{if } x' = x - 1,\ x \geq 1 \end{cases}$$

while the environment component Y is a Markov process with transition rates

$$r(1, 2) = \alpha_1 \quad \text{and} \quad r(2, 1) = \alpha_2.$$

Since the transition rates of (X, Y) have the form (3) and $\Phi(\cdot)$ is constant, Theorem 2 implies that $\pi(x, y) = \pi_X(x)\pi_Y(y)$, where

$$\pi_X(x) = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^x \quad \text{and} \quad \pi_Y(y) = 1 - \frac{\alpha_y}{\alpha_1 + \alpha_2}$$

and the form of $\pi_Y(y)$ is due to the fact that $Y(t)$ is an alternating renewal process. Yechiali and Naor [21] obtained this distribution assuming that $q(x, x+1; y)/q(x, x-1; y)$ is constant for each y. Yechiali [20] generalized this model for multiple environment states.

Example 8 *An M/M/1 Queue with Variable Capacity.*
We now consider a queueing system with variable capacity. Specifically, we assume Poisson arrivals and i.i.d exponentially distributed service times with rates that depend on the capacity of the system. We also assume that this capacity changes over time according to a birth-and-death process with rates that depend on the number of customers in the system. When the system is full, all arrivals are lost. We finally assume that the system capacity cannot decrease while the system is full.

We model the above system by the MIP (X, Y) where X denotes the number of customers in the system and $Y \geq 1$ denotes its capacity. Note that we always have $0 \leq X \leq Y$. For fixed y, the rates for X are

$$q(x, x'; y) = \begin{cases} \lambda\gamma(y) & x' = x + 1 \quad 0 \leq x \leq y - 1 \\ \mu\gamma(y) & x' = x - 1 \quad 1 \leq x \leq y \\ 0 & \text{otherwise} \end{cases}$$

while, for fixed x, the rates for Y are

$$r(y, y'; x) = \begin{cases} \alpha\rho(x) & y' = y + 1 \quad 0 \leq x \leq y \\ \beta\rho(x) & y' = y - 1 \quad 0 \leq x < y \\ 0 & \text{otherwise,} \end{cases}$$

where $\gamma(\cdot)$ and $\rho(\cdot)$ are positive functions.

Assume that (X, Y) is ergodic and let $\pi(x, y)$ be its stationary distribution. The following corollary follows from Theorem 2 and the fact that $\tilde{X}$ and $\tilde{Y}$ are birth-and-death processes.

Corollary 9 *The stationary distribution of* (X, Y) *exists if and only if* $\alpha < \beta$ *and* $\alpha\lambda < \beta\mu$ *and is given by*

$$\pi(x, y) = \left(\frac{\lambda}{\mu}\right)^x \left(\frac{\alpha}{\beta}\right)^{y-1} \pi(0, 1) \tag{6}$$

and

$$\pi(0, 1) = \left\{1 + \frac{\lambda}{\mu} + \frac{1}{1 - \lambda/\mu}\left[\frac{\alpha/\beta}{1 - \alpha/\beta} - \frac{\lambda^2}{\mu^2}\frac{\alpha\lambda/(\beta\mu)}{1 - \alpha\lambda/(\beta\mu)}\right]\right\}^{-1}. \tag{7}$$

Proof Using the equations

$$\pi(x, y) = \frac{\lambda}{\mu}\pi(x - 1, y) = \frac{\alpha}{\beta}\pi(x, y - 1).$$

and $\sum_{y=1}^{\infty}\sum_{x=0}^{y}\pi(x, y) = 1$, we have

$$\pi(0, 1)\left[1 + \frac{\lambda}{\mu} + \sum_{y=2}^{\infty}\left(\frac{\alpha}{\beta}\right)^{y-1}\sum_{x=0}^{y}\left(\frac{\lambda}{\mu}\right)^x\right] = 1,$$

and after some algebra

$$\pi(0, 1)\left\{1 + \frac{\lambda}{\mu} + \frac{1}{1 - \lambda/\mu}\left[\sum_{y=2}^{\infty}\left(\frac{\alpha}{\beta}\right)^{y-1} - \frac{\lambda^2}{\mu^2}\sum_{y=2}^{\infty}\left(\frac{\alpha\lambda}{\beta\mu}\right)^{y-1}\right]\right\} = 1.$$

Note here that $\pi(0, 1)$ exists if both summations in the last equation are finite. This is true if $\alpha < \beta$ and $\alpha\lambda < \beta\mu$. Computation of the two geometric series yields $\pi(0, 1)$ in (7). Finally, observe that the distribution $\pi(x, y)$ is given in terms of $\pi(0, 1)$ and therefore it exists only if $\pi(0, 1)$ does. □

3 Approximations for MIP's

The previous section considered MIP's with product form stationary distribution. Unfortunately, MIP's are quite complex and product form stationary distributions are quite difficult to obtain. This section proposes approximations for their stationary distributions.

The approximations in sections 3.1 and 3.2 use the concept of *nearly decomposable* and *nearly completely decomposable* matrices devised by Simon and Ando [19], and Ando and Fisher [1]. These matrices can be arranged to obtain a block diagonal structure where the elements within these blocks are larger in magnitude than the elements that are outside. Courtois [3] applied these concepts to stochastic matrices where the state space of the

corresponding Markov chains is partitioned into aggregate groups and the stationary probability of a particular state can be approximated by the stationary probability of this state within the corresponding aggregate times the probability of this aggregate in the macro process of the aggregates.

We begin with a summary of few results from the theory of decomposability for stochastic matrices, and then show how these results can be extended to rate matrices by using a uniformization technique.

Definition 10 *A* completely decomposable *matrix is a square matrix such that an identical permutation of rows and columns leaves a set of square submatrices on the principal diagonal and zeros elsewhere; if the resulting matrix has zeros everywhere below the principal submatrices but not necessarily above, then the original matrix is called* decomposable. Nearly completely decomposable *and* nearly decomposable *matrices are defined by replacing the zeros in the respective definitions by small nonzero numbers.*

A nearly completely decomposable stochastic matrix $\mathbf{P}$ can be written as

$$\mathbf{P} = \mathbf{P}^* + \epsilon \mathbf{C}, \tag{8}$$

where the blocks of the completely decomposable matrix

$$\mathbf{P}^* = \mathit{diagonal}\,(\mathbf{P}_1^*, \mathbf{P}_2^*, \ldots, \mathbf{P}_m^*)$$

are stochastic matrices and the scalar $\epsilon > 0$ is chosen as follows: For $u = 1, \ldots, m$ let $n(u)$ be the number of rows of $\mathbf{P}_u^*$ and let $p(i,u;j,v)$ denote the element of $\mathbf{P}$ at the intersection of the ith row of block u and the jth column of block v. Then

$$\epsilon = \max_{(i,u)} \left(\sum_{v \neq u} \sum_{j=1}^{n(v)} c(i,u;j,v) \right) \tag{9}$$

so that the elements of $\mathbf{C}$ satisfy $|c(i.u;j,v)| \leq 1$ and its rows sum to zero. We often call ϵ *the maximum degree of coupling between the submatrices* $\mathbf{P}_u^*$.

The following lemma summarizes the results in section 2.1 of Courtois [3].

Lemma 11 *Suppose that the $m \times m$ stochastic matrix with elements*

$$\overline{p}(u,v) = \sum_{i=1}^{n(u)} \pi(i;u) \sum_{j=1}^{n(v)} p(i,u;j,v) \tag{10}$$

is ergodic and let ν_u denote its stationary distribution. Then, for each $\delta > 0$, there is an $\epsilon_\delta > 0$ such that, for all $\epsilon < \epsilon_\delta$,

$$\max_{(i,u)} \mid \pi(i,u) - \pi(i;u)\nu_u \mid < \delta \tag{11}$$

for $i = 1, \ldots, n(u)$ and $u = 1, \ldots, m$. That is, $\pi(i,u) = \pi(i;u)\nu_u + O(\epsilon)$. □

The results of decomposability theory for stochastic matrices can be extended to Markov rate matrices by using the following uniformization result.

Lemma 12 *Assume* $\mathbf{\Lambda}$ *is a rate matrix for an ergodic Markov process, and let* $b = \sup_j \left(\sum_{k \neq j} \lambda_{jk}\right)$. *Then* $\mathbf{P} = \frac{1}{b}\mathbf{\Lambda} + \mathbf{I}$ *is a stochastic matrix and has the same stationary distribution as* $\mathbf{\Lambda}$.

Proof Clearly, $\mathbf{P}$ has non-negative elements and is stochastic since $\mathbf{P1} = \frac{1}{b}\mathbf{\Lambda 1} + \mathbf{I1} = \mathbf{1}$, where $\mathbf{1}$ is the column vector with unit elements. The second equality results from the fact that $\mathbf{\Lambda}$ is a rate matrix. Now if the row vector $\boldsymbol{\pi}$ is the stationary distribution of $\mathbf{\Lambda}$, then it is also the stationary distribution of $\mathbf{P}$ since $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}\left(\frac{1}{b}\mathbf{\Lambda} + \mathbf{I}\right) = \boldsymbol{\pi}$. □

3.1 Bivariate MIP's with Sluggish and Frequent Environment Changes

Consider a bivariate MIP (X, Y) with finite state space, transition rates given by (2), and stationary distribution $\pi(x, y)$, and denote the state spaces of X and Y by $\{1, \ldots, n\}$ and $\{1, \ldots, m\}$ respectively. We assume that one component, say X, has very large transition rates relative to the other "sluggish" component. The following theorem considers the first case and uses the results of the previous section to approximate the stationary distribution of (X, Y).

Theorem 13 *Assume that the rates* $q(x, x'; y)$ *are very large relative to* $r(y, y'; x)$. *Suppose that for each* y *the Markov process with rates* $q(x, x'; y)$ *is ergodic and let* $\pi(x; y)$ *denote its stationary distribution. Further, assume that the Markov process* $\overline{Y}$ *with the "averaged" rates*

$$\overline{r}(y, y') = \sum_x \pi(x; y) r(y, y'; x) \tag{12}$$

is ergodic and denote its stationary distribution by $\pi_{\overline{Y}}(y)$. *Let*

$$b = \max_{(x,y)} \left(\sum_{x'} q(x, x'; y) + \sum_{y'} r(y, y'; x) \right). \tag{13}$$

Then, for each $\delta > 0$, *there exists an* ϵ_δ *such that for* $b^{-1} < \epsilon_\delta$,

$$\max_{(x,y)} \mid \pi(x, y) - \pi(x; y)\pi_{\overline{Y}}(y) \mid < \delta. \tag{14}$$

Hence,

$$\pi(x, y) = \pi(x; y)\pi_{\overline{Y}}(y) + O(1/b). \tag{15}$$

Proof The rate matrix of the process (X, Y) can be written as

$$\mathbf{\Lambda} = \begin{pmatrix} (\mathbf{Q}_1 - \sum_{y'} \mathbf{R}_{1y'}) & \mathbf{R}_{12} & \cdots & \mathbf{R}_{1m} \\ \mathbf{R}_{21} & (\mathbf{Q}_2 - \sum_{y'} \mathbf{R}_{2y'}) & \cdots & \mathbf{R}_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{R}_{m1} & \mathbf{R}_{m2} & \cdots & (\mathbf{Q}_m - \sum_{y'} \mathbf{R}_{my'}) \end{pmatrix},$$

where the submatrix $\mathbf{Q}_y$ is defined by

$$\begin{aligned} (\mathbf{Q}_y)_{xx'} &= q(x, x'; y) \qquad \text{for } x \neq x' \\ (\mathbf{Q}_y)_{xx} &= -\sum_{x'} q(x, x'; y), \end{aligned}$$

and $\mathbf{R}_{yy'}$ is a diagonal matrix with $(\mathbf{R}_{yy'})_{xx} = r(y, y'; x)$.

Write

$$\mathbf{\Lambda} = \textit{diagonal}\,(\mathbf{Q}_1, \mathbf{Q}_2, \ldots, \mathbf{Q}_m) + \begin{pmatrix} -\sum_{y'} \mathbf{R}_{1y'} & \mathbf{R}_{12} & \cdots & \mathbf{R}_{1m} \\ \mathbf{R}_{21} & -\sum_{y'} \mathbf{R}_{2y'} & \cdots & \mathbf{R}_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{R}_{m1} & \mathbf{R}_{m2} & \ldots & -\sum_{y'} \mathbf{R}_{my'} \end{pmatrix},$$

or, $\mathbf{\Lambda} = \mathbf{Q} + \mathbf{R}$, and consider the stochastic matrix $\mathbf{P} = \frac{1}{b}\mathbf{\Lambda} + \mathbf{I}$, where b is defined in (13). Then $\mathbf{P}$ can be written as

$$\mathbf{P} = \frac{1}{b}(\mathbf{Q} + \mathrm{R}) + \mathbf{I} = \mathbf{P}^* + \frac{1}{b}\mathbf{R}, \tag{16}$$

where $\mathbf{P}^*$ consists of the m block diagonal submatrices $\mathbf{P}_y^* = \frac{1}{b}\ \mathbf{Q}_y + \mathbf{I}$. Lemma 12 implies that $\mathbf{P}^*$ is stochastic and has the same stationary distribution as $\mathbf{Q}$. Similarly, each submatrix $\mathbf{P}_y^*$ is stochastic with the same stationary distribution as $\mathbf{Q}_y$.

Note that $\mathbf{P}$ is nearly completely decomposable when b is large and has the form (8) with $\mathbf{C} = \mathbf{R}$ and $\epsilon = 1/b$ satisfying (9). From Lemma 11, if we let ν_y be the stationary distribution of the stochastic matrix with the averaged transition probabilities

$$\overline{p}_{yy'} = \sum_x \pi(x{:}\, y) \sum_{x'} p(x, y; x', y'),$$

where $p(x, y; x', y')$ are the elements of $\mathbf{P}$. then for any $\delta > 0$, there exists an ϵ_δ such that for all $\epsilon < \epsilon_\delta$,

$$\max_{(x,y)} \mid \pi(x, y) - \pi(x; y)\nu_y \mid < \delta.$$

It remains to show that $\nu_y = \pi_{\overline{Y}}(y)$. One has

$$p(x,y;x',y') = \begin{cases} \frac{1}{b}q(x,x';y) & x \neq x',\ y = y' \\ \frac{1}{b}r(y,y';x) & x = x',\ y \neq y' \\ 1 - \frac{1}{b}\big[\sum_u q(x,u;y) & \\ \qquad + \sum_v r(y,v;x)\big] & x = x',\ y = y' \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, for $y' = y$

$$\begin{aligned} \overline{p}_{yy} &= \sum_x \pi(x;y)\big[p(x,y;x,y) + \sum_{x' \neq x} p(x,y;x',y)\big] \\ &= 1 - \frac{1}{b}\sum_{y'} \overline{r}(y,y') \end{aligned}$$

and for $y' \neq y$

$$\overline{p}_{yy'} = \sum_x \pi(x;y)p(x,y;x,y') = \frac{1}{b}\,\overline{r}(y,y').$$

Lemma 12 now implies that ν_y satisfy the balance equations of $\overline{r}(y,y')$, and therefore $\nu_y = \pi_{\overline{Y}}(y)$. This proves (14). □

Remark 14 An advantage of the latter approximation is that it requires to solve $m+1$ systems of equations each with dimensions $n \times n$ instead of solving a system of size $(mn) \times (mn)$. It turns out (see Kant [7]) that since all submatrices $\mathbf{Q}_y$ have the same size, then the savings in such computations are maximized.

3.2 Approximations for Multivariate MIP's

In this section we use the results from Courtois [3] related to multilevel decomposability of shochastic matrices to generalize the approximations in Section 3.1 for multivariate MIP's.

We assume that the components of the MIP $\boldsymbol{X} = (X_1, \ldots, X_m)$ are already arranged so that

$$\lambda_1\left(x_1, x_1'; \boldsymbol{x}_1^c\right) >> \lambda_2\left(x_2, x_2'; \boldsymbol{x}_2^c\right) >> \cdots >> \lambda_m\left(x_m, x_m'; \boldsymbol{x}_m^c\right). \tag{17}$$

For simplicity, we consider a three-component MIP $\boldsymbol{X} = (X_1, X_2, X_3)$ and assume that X_j has n_j states. The proposed approach can be similarly generalized for MIP's with more components.

The rate matrix of $\boldsymbol{X}$ can be written as

$$\mathbf{\Lambda} = \textit{diagonal}\left(\mathbf{Q}^{(1)}(1,1), \mathbf{Q}^{(1)}(1,2), \ldots, \mathbf{Q}^{(1)}(n_2,n_3)\right) + \mathbf{R}^{(1)},$$

where for every (x_2, x_3), $\mathbf{Q}^{(1)}(x_2,x_3)$ is an $n_1 \times n_1$ rate matrix with elements $\lambda_1(x_1, x_1'; x_2, x_3)$.

Inequalities (17) suggest that $\mathbf{\Lambda}$ can be treated as nearly completely decomposable. Using Lemma 12, we uniformize $\mathbf{\Lambda}$ into a stochastic matrix

$$\mathbf{P}^{(1)} = \frac{1}{b_1}\mathbf{\Lambda} + \mathbf{I} = \mathbf{P}^{*(1)} + \frac{1}{b_1}\,\mathbf{R}^{(1)},$$

where

$$b_1 = max_{(x_1,x_2,x_3)}\left\{\lambda_1(x_1,x_1';x_2,x_3) + \lambda_2(x_2,x_2';x_1,x_3) + \lambda_3(x_3,x_3';x_1,x_2)\right\}$$

and the matrix $\mathbf{P}^{*(1)}$ is completely decomposable with n_2n_3 blocks given for every (x_2,x_3) by

$$\mathbf{P}^{*(1)}(x_2,x_3) = \frac{1}{b_1}\mathbf{Q}^{(1)}(x_2,x_3) + \mathbf{I}.$$

Assume that the Markov process $(X_1; x_2, x_3)$ corresponding to the submatrix $\mathbf{Q}^{(1)}(x_2,x_3)$ is ergodic for every (x_2,x_3) and let $\pi^{(1)}(x_1;x_2,x_3)$ be its stationary distribution. Then we use these distributions to obtain the averaged rates

$$\begin{aligned}
\overline{\lambda}_2^{(1)}(x_2,x_2';x_3) &= \sum_{x_1}\pi^{(1)}(x_1;x_2,x_3)\lambda_2(x_2,x_2';x_1,x_3)\\
\overline{\lambda}_3^{(1)}(x_3,x_3';x_2) &= \sum_{x_1}\pi^{(1)}(x_1;x_2,x_3)\lambda_3(x_3,x_3';x_1,x_2).
\end{aligned}$$

These rates define a new Markov process, say $(\overline{X}_2^{(1)}, \overline{X}_3^{(1)})$, in the Cartesian product of the state spaces of X_2 and X_3. Observe that if we let $\pi^{(2)}(x_2,x_3)$ be the joint stationary distribution of $(\overline{X}_2^{(1)}, \overline{X}_3^{(1)})$, then Theorem 13 implies

$$\pi(x_1,x_2,x_3) = \pi^{(1)}(x_1;x_2,x_3)\pi^{(2)}(x_2,x_3) + O(\frac{1}{b_1}).$$

Moreover, since we assumed that $\lambda_2(x_2,x_2';x_1,x_3) >> \lambda_3(x_3,x_3';x_1,x_2)$, then it is clear that

$$\overline{\lambda}_2^{(1)}(x_2,x_2';x_3) >> \overline{\lambda}_3^{(1)}(x_3,x_3';x_2). \tag{18}$$

If we now let $\overline{\mathbf{\Lambda}}^{(1)}$ be the $(n_2n_3)\times(n_2n_3)$ rate matrix of $(\overline{X}_2^{(1)}, \overline{X}_3^{(1)})$, then we can arrange $\overline{\mathbf{\Lambda}}^{(1)}$ so that it will have principal diagonal blocks with larger elements than those that are outside these blocks. Specifically, we write

$$\overline{\mathbf{\Lambda}}^{(1)} = \textit{diagonal}\left(\mathbf{Q}^{(2)}(1), \mathbf{Q}^{(2)}(2), \ldots, \mathbf{Q}^{(2)}(n_3)\right) + \mathbf{R}^{(2)},$$

where $\mathbf{Q}^{(2)}(x_3)$ is an $n_2 \times n_2$ rate matrix with elements $\overline{\lambda}_2^{(1)}(x_2, x_2'; x_3)$. We treat $\overline{\mathbf{\Lambda}}^{(1)}$ as nearly completely decomposable and repeat the above procedure.

Uniformizing $\overline{\mathbf{\Lambda}}^{(1)}$, we obtain

$$\mathbf{P}^{(2)} = \frac{1}{b_2}\overline{\mathbf{\Lambda}}^{(1)} + \mathbf{I} = \mathbf{P}^{*(2)} + \frac{1}{b_2}\mathbf{R}^{(2)},$$

where

$$b_2 = max_{(x_2,x_3)}\left\{\overline{\lambda}_2^{(1)}(x_2, x_2'; x_3) + \overline{\lambda}_3^{(1)}(x_3, x_3'; x_2)\right\}$$

and the matrix $\mathbf{P}^{*(2)}$ is completely decomposable with n_3 blocks given for every x_3 by

$$\mathbf{P}^{*(2)}(x_3) = \frac{1}{b_2}\mathbf{Q}^{(2)}(x_3) + \mathbf{I}.$$

Assume that, for each x_3, the Markov process $(\overline{X}_2^{(1)}; x_3)$ with rate matrix $\mathbf{Q}^{(2)}(x_3)$ is ergodic, and denote its stationary distribution by $\pi^{(2)}(x_2; x_3)$. We then average $\overline{\lambda}_3^{(1)}(x_3, x_3'; x_2)$ by this distribution to obtain the rates

$$\overline{\lambda}^{(2)}(x_3, x_3') = \sum_{x_2} \pi^{(2)}(x_2; x_3)\overline{\lambda}_3^{(1)}(x_3, x_3'; x_2).$$

We finally assume that these rates define a new ergodic process $\overline{X}_3^{(2)}$ on the state space of X_3 with stationary distribution $\pi^{(3)}(x_3)$. Theorem 13 now implies

$$\pi^{(2)}(x_2, x_3) = \pi^{(2)}(x_2; x_3)\pi^{(3)}(x_3) + O(\frac{1}{b_2}).$$

At this point we stop and write the stationary distribution of the process $\boldsymbol{X}$ as

$$\pi(x_1, x_2, x_3) = \pi^{(1)}(x_1; x_2, x_3)\pi^{(2)}(x_2; x_3)\pi^{(3)}(x_3) + O(\frac{1}{b_1} + \frac{1}{b_2}).$$

3.3 Bivariate MIP's with Dominating Environment

We now consider the case in which the process (X, Y) spends a large fraction of time in a particular environment while its visits to the remaining environments are temporary. Without loss of generality, we assume that the environment Y has only two states, say 1 and 2, with the second being the dominant state. Theorem 15 states that the process (X, Y) can be approximated by the process that restricts the system component X to environment 2.

Theorem 15 *Assume that the state space of Y has only two states where environemt 2 dominates environment 1, that is, for all x, $r(2,1;x) \to 0$ while $r(1,2;x)$ are bounded away from zero. Then the stationary distribution of (X,Y) satisfies*

$$\pi(x,y) \to \begin{cases} 0 & \text{if } y = 1 \\ \pi_X(x;2) & \text{if } y = 2, \end{cases} \tag{19}$$

where $\pi_X(x;2)$ is the stationary probability distribution of the process with rates $q(x,x';2)$.

Proof As in section 3.1, the rate matrix of (X,Y) can be written as

$$\mathbf{\Lambda} = \begin{pmatrix} \mathbf{Q}_1 - \mathbf{R}_{12} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{Q}_2 - \mathbf{R}_{21} \end{pmatrix},$$

where $\mathbf{R}_{21} \to \mathbf{0}$. Thus $\mathbf{\Lambda} \to \mathbf{\Lambda}^*$ where

$$\mathbf{\Lambda}^* = \begin{pmatrix} \mathbf{Q}_1 - \mathbf{R}_{12} & \mathbf{R}_{12} \\ \mathbf{0} & \mathbf{Q}_2 \end{pmatrix}$$

implying that $\pi(x,y)$ converges to the stationary distribution of $\mathbf{\Lambda}^*$. Since $\mathbf{R}_{12}$ is nonzero, then it is clear that the class $\{(x,1)\}$ of $\mathbf{\Lambda}^*$ is transient while the class $\{(x,2)\}$ is recurrent and (19) follows. □

3.4 MMP's with Many Non-dominating Environments

In this section, we discuss an ergodic Markov modulated process where the system component is controlled by an environment that is almost equally likely to be in any one of a large number of states.

Assume that (X,Y) is an MMP with transition rates

$$\lambda(x,y;x',y') = \begin{cases} q(x,x';y) & x \neq x', y = y' \\ r(y,y') & x = x', y \neq y' \\ 0 & \text{otherwise.} \end{cases}$$

It is easy to see that the marginal stationary distribution $\pi_Y(y)$ of Y is the solution to the balance equations of the rates $r(y,y')$. Let m be the number of states of Y. We say that (X,Y) has *non-dominating* environments if $\pi_Y(y) \to 0$ as $m \to \infty$ for each y.

The following theorem proposes an approximation of the stationary distribution $\pi(x,y)$ for large m.

Theorem 16 *Suppose that the rates $q(x,x';y)$ are bounded for all x,x' and y. Also, assume that the process $\overline{X}$ with rates*

$$\overline{q}(x,x') = \sum_y \pi_Y(y) q(x,x';y) \qquad \textit{for all } x,x'$$

is ergodic with stationary distribution $\pi_{\overline{X}}(x)$. Then the stationary distribution of the MMP (X,Y) with m non-dominating environment states satisfies

$$\pi(x,y) - \pi_{\overline{X}}(x)\pi_Y(y) \to 0 \qquad \textit{as } m \to \infty. \tag{20}$$

Proof The balance equations of (X,Y) are

$$\begin{aligned} -\pi(x,y)\left[\sum_{x'} q(x,x';y) + \sum_{y'} r(y,y')\right] \\ + \sum_{x'} \pi(x',y) q(x',x;y) + \sum_{y'} \pi(x,y') r(y',y) = 0. \end{aligned} \tag{21}$$

Let $\epsilon(x,y) = \pi(x,y) - \pi_{\overline{X}}(x)\pi_Y(y)$ and note that the balance equations of the process $\overline{X}$ imply $\sum_x \sum_y \epsilon(x,y) = 0$.

If we replace $\pi(x,y)$ by $\pi_{\overline{X}}(x)\pi_Y(y)$ in the balance equations (21), we obtain the balance residuals given by

$$\begin{aligned} \delta(x,y) &= \pi_{\overline{X}}(x)\pi_Y(y)\left[\sum_{x'} q(x,x';y) + \sum_{y'} r(y,y')\right] \\ &\quad - \sum_{x'} \pi_{\overline{X}}(x')\pi_Y(y) q(x',x;y) - \sum_{y'} \pi_{\overline{X}}(x)\pi_Y(y') r(y',y). \end{aligned} \tag{22}$$

The balance equations for Y simplify $\delta(x,y)$ to

$$\delta(x.y) = \pi_Y(y)\left[\pi_{\overline{X}}(x)\sum_{x'} q(x,x';y) - \sum_{x'} \pi_{\overline{X}}(x') q(x',x;y)\right]. \tag{23}$$

Now we can relate $\epsilon(x,y)$ to $\delta(x,y)$ by subtracting (21) from (22) to obtain

$$\begin{aligned} \delta(x,y) &= -\epsilon(x,y)\left[\sum_{x'} q(x,x';y) + \sum_{y'} r(y,y')\right] \\ &\quad + \sum_{x'} \epsilon(x',y) q(x',x;y) + \sum_{y'} \epsilon(x,y') r(y',y), \end{aligned}$$

or in matrix form $\boldsymbol{\Lambda\epsilon} = \boldsymbol{\delta}$, where $\boldsymbol{\Lambda}$ is the rate matrix of the MMP. Since $rank(\boldsymbol{\Lambda}) = mn - 1$, we use the equation $\sum_x \sum_y \epsilon(x,y) = 0$ to replace one row of $\boldsymbol{\Lambda}$ by ones and

the corresponding $\delta(x,y)$ by 0. The resulting matrix $\mathbf{\Lambda}_a$ is nonsingular and the system $\mathbf{\Lambda}_a\boldsymbol{\epsilon} = \boldsymbol{\delta}_a$ has solution $\boldsymbol{\epsilon} = \mathbf{\Lambda}_a^{-1}\boldsymbol{\delta}_a$.

Since the rates $q(x,x';y)$ are bounded and $\pi_Y(y) \to 0$ as $m \to \infty$, equation (23) implies $\delta(x,y) \to 0$ and, consequently, $\epsilon(x,y) \to 0$ for all x and y. □

Remark 17 The marginal distribution of the process X is approximated by $\pi_X(x) \approx \pi_{\overline{X}}(x)$.

3.5 Approximating Functionals of Bivariate MIP's

Functionals of ergodic Markov processes are used in evaluating long-run performance measures such as marginal distributions, expected queue lengths, throughput rates and average costs.

Let $g(x,y)$ be the cost rate when the bivariate MIP (X,Y) is at state (x,y), and let $h(x,y;x',y')$ be the cost of a transition from state (x,y) to (x',y'). Ergodic theorems for Markov processes imply, with probability one, that the long-run average costs associated with g and h are

$$\lim_{t\to+\infty} \frac{1}{t}\int_0^t g\big(X(s),Y(s)\big)ds = \sum_x \sum_y \pi(x,y)\ g(x,y) \equiv E\big[g(X,Y)\big]$$

and

$$\begin{aligned}
&\lim_{t\to+\infty} \frac{1}{t}\sum_{s\le t} h\big(X(s-),Y(s-);X(s),Y(s)\big) \\
&= \sum_x \sum_y \pi(x,y) \left[\sum_{x'}\sum_{y'} \lambda(x,y;x',y')\ h(x,y;x',y')\right] \\
&\equiv E\big[\lambda(X,Y;X',Y')h(X,Y;X',Y')\big].
\end{aligned}$$

In the case of sluggish environment changes, the approximation in Theorem 13 yields

$$\begin{aligned}
E\big[g(X,Y)\big] &\approx \sum_x \sum_y \pi(x;y)\pi_{\overline{Y}}(y)\ g(x,y) \\
&\approx \sum_y \pi_{\overline{Y}}(y) \sum_x \pi(x;y)g(x,y).
\end{aligned}$$

On the other hand. since the transition rates for an MIP can be written as

$$\lambda(x,y;x',y') = q(x,x';y)1(y'=y) + r(y,y';x)1(x'=x),$$

the expectation $E[\lambda(X,Y;X',Y')h(X,Y;X',Y')]$ can be approximated by

$$\sum_y \pi_{\overline{Y}}(y) \sum_x \pi(x;y) \left[\sum_{x'} q(x,x';y)h(x,y;x',y) + \sum_{y'} r(y,y';x)h(x,y;x,y') \right].$$

Expressions for MIP's with frequent environment transitions and MMP's with non-dominating environments are obtained analogously.

4 Illustrations of MIP's

In this section, we consider several systems that can be modelled as MIP's and give exact or approximate expressions for their stationary distributions.

4.1 M/M/Y Queueing Systems with Random Number of Servers

We first consider a queueing system where the number of servers varies with time. Such variations in the number of servers may result by assigning more servers when the queue builds up or by using the servers elsewhere when the system becomes relatively uncongested.

The system under study is modelled by a process (X,Y), where X counts the number of customers and Y counts the number of servers. When y servers are attending, arrivals occur according to a Poisson process with rate $\lambda(x;y)$ and the service rate is $\mu(x;y)$. Also, when the system contains x customers, the number of servers fluctuates according to a Markov process with rates $r(y,y';x)$.

Whenever a server is removed from the system while serving a customer, the customer rejoins the queue. If a new server becomes active, then it immediately starts service of one of the waiting customers or remains idle if the queue is empty. We assume that (X,Y) is ergodic and use the results obtained in the previous two sections to characterize the stationary distribution of this process.

Product Form Stationary Distribution
Suppose that the transition rates of (X,Y) are

$$q(x,x';y) = \begin{cases} \lambda(x)\gamma(y) & x' = x+1 \\ \mu(x)\gamma(y) & x' = x-1 \\ 0 & \text{otherwise} \end{cases}$$

and

$$r(y,y';x) = r(y,y')\rho(x),$$

where y takes on positive values and the functions λ, μ and γ are non-negative functions with $\mu(0) = 0$. Then by Theorem 2, the stationary distribution has the product form $\pi(x,y) = \pi_X(x)\pi_Y(y)$, where $\pi_X(x)$ is the stationary distribution of the $M/M/1$ process defined on the state space of X with arrival and departure rates $\lambda(x)$ and $\mu(x)$ respectively, and $\pi_Y(y)$ is the stationary distribution of the Markov process defined on the state space of Y with rates $r(y,y')$.

Approximate Product Form Stationary Distribution

Assume that the state spaces for X and Y are both finite and consider the case where the number of servers changes at a slow rate. For example, it may take some time to hire a new server and make him active in the system, or it may take a while to find a new assignment for an existing server before ceasing its service. Formally, we assume that the rates $r(y,y')$ are small compared to $q(x,x';y)$. Using Theorem 13, we can approximate the stationary distribution of (X,Y) by $\pi(x,y) \approx \pi(x;y)\pi_{\overline{Y}}(y)$, where $\pi(x;y)$ is the stationary distribution of an $M/M/y$ queueing system with fixed number of servers equal to y, and $\pi_{\overline{Y}}(y)$ is the stationary distribution of a process $\overline{Y}$ defined by the averaged rates

$$\overline{r}(y,y') = \sum_x \pi_X(x;y) r(y,y';x).$$

Further, the expected number of customers in the system can be approximated as

$$E[X] \approx \sum_y \pi_{\overline{Y}}(y) \sum_x x\pi(x;y).$$

A Numerical Example

To assess the above approximation numerically, we consider a system with room for 10 customers and number of servers taking values in $\{1.2,3\}$. The arrival and departure rates in the following table depend only on the number of servers

y	$\lambda(y)$	$\mu(y)$
1	9	10
2	15	20
3	20	30

while the rates of jumps in the number of servers are listed in the table below.

	x		
	0-3	4-7	8-10
$r(1,2;x)$	0.05	0.15	0.20
$r(1,3;x)$	0.01	0.05	0.25
$r(2,1;x)$	0.15	0.05	0.05
$r(2,3;x)$	0.01	0.05	0.15
$r(3,1;x)$	0.20	0.25	0.01
$r(3,2;x)$	0.25	0.05	0.01

The following table lists the exact and approximate stationary distributions. The calculations were performed by using MATHCAD.

	$y=1$		$y=2$		$y=3$	
x	exact	approximate	exact	approximate	exact	approximate
0	0.0649	0.0642	0.1193	0.1203	0.0333	0.0333
1	0.0584	0.0587	0.0895	0.0902	0.0222	0.0222
2	0.0525	0.0520	0.0671	0.0677	0.0148	0.0148
3	0.0472	0.0468	0.0503	0.0507	0.0099	0.0099
4	0.0425	0.0421	0.0378	0.0381	0.0066	0.0066
5	0.0382	0.0379	0.0283	0.0285	0.0044	0.0044
6	0.0343	0.0341	0.0213	0.0214	0.0029	0.0029
7	0.0308	0.0307	0.0160	0.0161	0.0020	0.0019
8	0.0277	0.0276	0.0112	0.0120	0.0013	0.0013
9	0.0249	0.0249	0.0090	0.0090	0.0090	0.0009
10	0.0224	0.0224	0.0067	0.0068	0.0006	0.0006

The largest absolute difference of the two distributions is 0.0019 while the largest relative difference is 2.416%. Also, the exact and approximate marginal stationary distributions for the number of customers in the system are

x	exact	approximate
0	0.2175	0.2177
1	0.1701	0.1702
2	0.1344	0.1344
3	0.1075	0.1074
4	0.0868	0.0868
5	0.0709	0.0708
6	0.0585	0.0584
7	0.0488	0.0487
8	0.0410	0.0410
9	0.0348	0.0348
10	0.0298	0.0297

while the exact and approximate mean queue lengths are 3.0943 and 3.0923, respectively.

4.2 M/M/1 Queueing Systems with Two Interacting Types of Customers

In this section we consider an $M/M/1$ queueing system used by two types of customers having different arrival and service rates and such that these rates are larger for one type than the other. That is, the customers of the type with larger rates arrive more frequently and require smaller service times in the system. We assume that this system has service sharing discipline and that the rates for one type depend on the number of customers of the other type present in the system.

This system can be modelled by a bivariate MIP (X_1, X_2) where X_i counts the number of customers of type i who arrive according to a state-dependent Poisson process with rates $\lambda_i(x_1, x_2)$ and have i.i.d exponentially distributed service times with rates $\mu_i(x_1, x_2)$. Thus, the transition rates of (X_1, X_2) are

$$q_1(x_1, x_1'; x_2) = \begin{cases} \lambda_1(x_1, x_2) & x_1' = x_1 + 1 \\ \mu_1(x_1, x_2) & x_1' = x_1 - 1 \\ 0 & \text{otherwise} \end{cases}$$

and

$$q_2(x_2, x_2'; x_1) = \begin{cases} \lambda_2(x_1, x_2) & x_2' = x_2 + 1 \\ \mu_2(x_1, x_2) & x_2' = x_2 - 1 \\ 0 & \text{otherwise.} \end{cases}$$

A product form solution for the balance equations of this system is very hard to obtain. We then use the results of section 3.1 to approximate its stationary distribution. To this end, we assume that the system has finite capacity and all arrivals to a full system are turned away.

Let

$$b = max_{x_1, x_2} \{\lambda_1(x_1, x_2) + \mu_1(x_1, x_2) + \lambda_2(x_1, x_2) + \mu_2(x_1, x_2)\}$$

and assume that $q_1(x_1, x_1'; x_2)$ are large relatively to $q_2(x_2, x_2'; x_1)$. Then Theorem 13 implies

$$\pi(x_1, x_2) = \pi(x_1; x_2)\pi_{\overline{X}_2}(x_2) + O\left(\frac{1}{b}\right),$$

where $\pi_{X_1}(x_1; x_2)$ is the stationary distribution of a Markov process $(X_1; x_2)$ with rates $\lambda_1(x_1, x_2)$ and $\mu_1(x_1, x_2)$ for fixed x_2, and $\pi_{\overline{X}_2}(x_2)$ is the stationary distribution of a Markov process $\overline{X}_2$ with the averaged rates $\overline{\lambda}_2(x_2) = \sum_{x_1} \pi_{X_1}(x_1; x_2)\lambda_2(x_1, x_2)$ and $\overline{\mu}_2(x_2) = \sum_{x_1} \pi_{X_1}(x_1; x_2)\mu_2(x_1, x_2)$.

4.3 Approximation for Markov Modulated M/M/1/K Systems with Many Non-dominating Environments

In this section, we illustrate the approximation in Theorem 16 for a stable $M/M/1$ queueing system with fixed finite capacity. We assume that the arrival and service rates change according to an extraneous birth-and-death process having a large number of almost equally likely states. We represent this model by the process (X, Y), where X is the system component with rates

$$q(x, x'; y) = \begin{cases} \lambda(y) & x' = x + 1 \\ \mu(y) & x' = x - 1 \\ 0 & \text{otherwise} \end{cases}$$

defined for a given y, and Y is the environment component with rates $r(y, y')$ that are independent of the system state.

Suppose that the system has capacity $K = 3$, and the environment has $m = 20$ states. We assume that the birth and death process Y has stationary distribution

y	1	2	3	4	5	6	7	8	9	10
$\pi_Y(y)$	0.048	0.048	0.048	0.048	0.050	0.049	0.048	0.049	0.049	0.049
y	11	12	13	14	15	16	17	18	19	20
$\pi_Y(y)$	0.049	0.050	0.049	0.049	0.051	0.052	0.053	0.053	0.053	0.054

The arrival and departure rates for X are given by

y	1	2	3	4	5	6	7	8	9	10
$\lambda(y)$	1.00	1.00	1.00	2.00	2.00	2.00	2.00	3.00	3.00	4.00
$\mu(y)$	1.5	1.75	1.25	2.50	2.75	3.00	5.00	4.00	3.25	5.00
y	11	12	13	14	15	16	17	18	19	20
$\lambda(y)$	4.00	4.00	4.00	4.00	5.00	5.00	5.00	5.00	6.00	6.00
$\mu(y)$	4.25	4.75	6.00	7.00	6.00	6.50	7.00	7.50	7.00	8.00

and, averaged by $\pi_Y(y)$, yield the arrival rate $\overline{\lambda} = 3.505$ and the service rate $\overline{\mu} = 4.768$.

The exact and approximate marginal stationary distributions for the system component X are listed below.

x	exact	approximate
0	0.381	0.374
1	0.274	0.275
2	0.199	0.202
3	0.146	0.149

Also, the exact and approximate mean queue lengths are, respectively, 1.11 and 1.125.

4.4 An Approximation for an M/M/1 Queue with Variable Capacity

Suppose that the process Y that counts the capacity of the queue in Example 8 has finite state space, say $\{1, 2, \ldots, m\}$, and $\alpha(x) = \alpha$, $\beta(x) = \beta$ for all x. We use the approximation in Theorem 16 to estimate the stationary distribution of the MMP (X, Y) when $\alpha \approx \beta$. Then $\pi_Y(y) \approx 1/m$ for $y = 1, 2, \ldots, m$ and, for large m, we have

$$\pi(x, y) \approx \frac{1}{m} \pi_{\overline{X}}(x),$$

where $\pi_{\overline{X}}(x)$ is the stationary distribution of the process $\overline{X}$ with rates

$$\overline{q}(x, x+1) = \sum_{y=x}^{m} \pi_Y(y)\lambda(y) \quad \text{and} \quad \overline{q}(x, x-1) = \sum_{y=x}^{m} \pi_Y(y)\mu(y).$$

4.5 Closed Queueing Networks with Multi-Mode Operating Nodes as Markov Modulated Processes

In this section, we study closed network processes with nodes that may have several service rates. We will first discuss a model that has a product form stationary distribution under different conditions than those of Theorem 2. We then use the results of Theorem 16 to derive an approximation to the stationary distribution of such network processes.

Consider a network with J nodes described by the pair (X, Y), where $X = (X_1, \ldots, X_J)$ is the system component with X_j counting the number of customers at node j, and $Y = (Y_1, \ldots, Y_J)$ is the environment component with Y_j denoting the service rate at node j. The network is closed with a fixed number of customers, say N. A typical state of X is $x = (x_1, \ldots, x_J)$, where $\sum_{j=1}^{J} x_j = N$. Each Y_j has m_j states so that the state space of Y is finite with cardinality $m = \prod_{j=1}^{J} m_j$.

A transition in the system component X occurs when a customer moves from one node to another. When (X, Y) is in state (x, y), the time to a potential movement of a customer from node j to node k is exponentially distributed with rate $q(x, T_{jk}x; y)$, for $x_j > 0$. Here $T_{jk}(x) = x - e_j + e_k$, where e_j is a $J \times 1$ vector of zeros except that its jth coordinate is equal to one. On the other hand, a transition in Y occurs when the service rate at a node j changes from y_j to y_j' independently of the service rates at the other nodes. We assume that these changes occur independently of X according to a Markov process with rates $r_j(y_j, y_j')$. We assume in addition that the transitions in X and Y occur only one-at-a-time so that the process (X, Y) is an MMP.

4.5.1 Closed Reversible Jackson Network with Nodes Subject to Breakdowns

In this case, nodes can be either up or down for exponentially distributed random times, independently from each other with mean time before a breakdown $1/\alpha_j$ and mean repair time $1/\beta_j$ for node j.

If we denote the up or down mode at the node j by $y_j = 1$ and 0 respectively, then we can write the transition rates of Y caused by the change of modes at node j as

$$r_j(y_j, 1 - y_j) = \theta_j(y_j) = \alpha_j y_j + \beta_j(1 - y_j).$$

The limiting distribution of the state of node j is given by

$$\pi_j(y_j) = \frac{\theta_j(1 - y_j)}{\theta_j(y_j) + \theta_j(1 - y_j)} \tag{24}$$

and satisfies

$$\pi_j(y_j)\theta_j(y_j) = \pi_j(1 - y_j)\theta_j(1 - y_j). \tag{25}$$

We assume that the arrival and service rates at a down node are zero. Also, the transition rates for the component X are given by

$$q(x, T_{jk}x; y) = \lambda_{jk}\phi_j(x_j)1(y_j = y_k = 1), \tag{26}$$

where λ_{jk} is the routing intensity from node j to node k and $\phi_j(x_j)$ is the service rate at node j when it contains x_j customers.

Now we are in a position to show that if the network has reversible routings as defined in condition (27) of the following theorem, then its stationary distribution has a product form.

Theorem 18 *Assume that there exist $w_1, w_2, \ldots, w_J$ such that for all pairs of nodes (j, k)*

$$w_j\lambda_{jk} = w_k\lambda_{kj}. \tag{27}$$

If the process (X, Y) defined above is ergodic, then it is reversible and has a product form stationary distribution given by

$$\pi(x, y) = \pi_X(x)\pi_Y(y) \tag{28}$$

where

$$\pi_X(x) = c\prod_{j=1}^{J} w_j^{x_j} \prod_{\nu=1}^{x_j} \phi_j^{-1}(\nu) \tag{29}$$

and

$$\pi_Y(y) = \prod_{j=1}^{J} \pi_j(y_j). \tag{30}$$

Proof Equations (27) and (29) yield

$$\pi_X(x)\lambda_{jk}\phi_j(x_j) = \pi_X(T_{jk}x)\lambda_{kj}\phi_k(x_k + 1) \tag{31}$$

implying the reversibility of X. On the other hand, (25) and (30) yield

$$\pi_Y(y)\theta_j(y_j) = \pi_Y(y_1, \ldots, y_{j-1}, 1 - y_j, y_{j+1}, \ldots, y_m)\theta_j(1 - y_j) \tag{32}$$

implying that Y is also reversible. The proof follows. □

4.5.2 Approximation in Closed Networks with Nodes having Many Non-dominating Multiple Service Modes.

Here, we will use the approximation for non-dominating environment processes described in section 3.4 for closed Jackson networks where each node may have more than one mode of operation.

We assume that the Markov process Y is ergodic and independent of X. Therefore the stationary distribution of Y is given by $\pi_Y(y) = \prod_{j=1}^{J} \pi_j(y_j)$, where $\pi_j(y_j)$ is given by (24). In addition, we assume that every node j has all its service modes almost equally likely, in the sense that $\pi_j(y_j) \approx 1/m_j$. This will imply $\pi_Y(y) \approx 1/m$. Hence, assuming the process $\overline{X}$ defined on the state space of X with the averaged rates $\overline{q}_{jk}(x) = \sum_y \pi_Y(y) q(x, T_{jk}x; y)$ is ergodic with stationary distribution $\pi_{\overline{X}}(x)$ and that m_j are large, we apply Theorem 16 to approximate the stationary distribution of (X, Y) by $\pi(x, y) \approx \frac{1}{m}\pi_{\overline{X}}(x)$ and the marginal distribution of X by $\pi_X(x) \approx \pi_{\overline{X}}(x)$.

References

[1] A. Ando and F. M. Fisher. Near-decomposability, partition and aggregation, and the relevance of stability discussions. *International Economic Review*, 4(1):53–67, 1963.

[2] D. Y. Burman and D. R. Smith. An asymptotic analysis of a queueing system with Markov-modulated arrivals. *Operations Research*, 34:105–119, 1986.

[3] P. J. Courtois. *Decomposability, Queueing and Computer System Applications.* Academic Press, 1977.

[4] M. Eisen and M. Tainitier. Stochastic variations in queueing processes. *Operations Research*, 11:922–927, 1963.

[5] H. Heffes and D. M. Lucantoni. A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance. *IEEE Journal on Selected Areas in Communications*, SAC-4:856–867, 1986.

[6] W. Henderson, C.E.M. Pearce, P.G. Taylor, and N.M. van Dijk. Closed queueing networks with batch services. *Queueing Systems*, 6:59–70, 1990.

[7] K. Kant. *Introduction to Computer System Performance Evaluation.* McGraw-Hill, Inc., 1991.

[8] C. Knessl, B.J. Matkowsky, Z. Schuss, and C. Tier. A Markov-modulated M/G/1 queue I: stationary distribution. *Queueing Systems*, 1:355–374, 1987.

[9] K.S. Meier-Hellstern. A fitting algorithm for Markov-modulated Poisson processes having two arrival rates. *European Journal of Operational Research*, 29:370–377, 1987.

[10] I. L. Mitrany and B. Avi-Itzhac. A many-server queue with service interruptions. *Operations Research*, 16:628–638, 1968.

[11] M. F. Neuts. A queue subject to extraneous phase changes. *Advanced Applied Probability*, 3:78–119, 1971.

[12] M. F. Neuts. Further results on the M/M/1 queue with randomly varying rates. *Operations Research*, 15(4):158–168, 1978.

[13] M. F. Neuts. The M/M/1 queue with randomly varying arrivals and service rates. *Operations Research*, 15(4):139–157, 1978.

[14] M. F. Neuts. *Matrix-Geometric Solutions in Stochastic Models, An Algorithmic Approach.* The John Hopkins University Press, 1981.

[15] M. F. Neuts and D. M. Lucantoni. A Markovian queue with N servers subject to breakdowns and repairs. *Management Science*, 25(9):849–861, 1979.

[16] N. U. Prabhu and Y. Zhu. Markov-modulated queueing systems. *Queueing Systems*, 5:215–246, 1989.

[17] P. Purdue. The M/M/1 queue in a Markovian environment. *Operations Research*, 22:562–569, 1974.

[18] G. J.K. Regterschot and H. A. de Smit. The Queue M/G/1 with Markov modulated arrivals and services. *Mathematics of Operations Research*, 11(3):465–483, 1986.

[19] H. A. Simon and A. Ando. Aggregation of variables in dynamic systems. *Econometrica*, 29(2):111–138, 1961.

[20] U. Yechiali. A queueing-type birth-and-death process defined on a continuous-time Markov chain. *Operations Research*, 21:604–609, 1973.

[21] U. Yechiali and P. Naor. Queuing problems with heterogeneous arrivals and service. *Operations Research*, 19:722–734, 1971.

[22] Y. Zhu. A Markov-modulated M/M/1 queue with group arrivals. *Queueing Systems*, 8:255–264, 1991.

[23] Y. Zhu and N. U. Prabhu. Markov-modulated PH/G/1 queueing systems. *Queueing Systems*, 9:313–322, 1991.

A Note on State Space Decomposition Methods for Computing Performance Measures of Stochastic Flow Networks

Christos Alexopoulos
School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332-0205

Key Words – Maximum flow, minimum cut, network reliability.

Reader Aids –

Purpose: Correction of existing algorithms.

Special math needed for explanations: Probability and statistics.

Special mathematics needed to use results: Same.

Results useful to: Reliability analysis.

Summary and Conclusions –

Consider a flow network with single source s and single sink t with demand $d > 0$. Assume that the nodes do not restrict flow transmission and the arcs have finite random discrete capacities. This paper has two objectives: (1) It corrects errors in well-known algorithms by Doulliez and Jamoulle (1972) for computing the probability that the demand is satisfied (or network reliability), the probability that an arc belongs to a minimum cut which limits the flow below d, and the probability that a cut limits the flow below d; (2) It discusses the applicability of these procedures.

This research was supported by the Air Force Office of Scientific Research under grant F49620-93-1-0043. Reproduction in whole or part is permitted for any purpose of the United States Government.

1. Introduction

A primary objective in the design of a transmission network is the establishment of flow between a source node and a sink node that exceeds a specified value d. If such a feasible flow does not exist, the determination of the locations of the bottlenecks that limit the flow of commodity is an important problem in network design. The solutions to these problems are well known in the case of deterministic capacities. However, the capacities are often random variables in practice and the probability that a feasible flow exists is a measure of system performance, often called the *network reliability.* Another measure is the probability that a set of arcs (cut) limits the value of maximum flow below d.

Specifically, let $G = (\mathcal{N}, \mathcal{A}, s, t)$ denote the flow network where $\mathcal{N}$ is the set of nodes, $\mathcal{A} = \{1,..., a\}$ is the set of arcs, s is the source, and t is the sink. Suppose that the nodes do not limit flow transmission and that the capacity of arc j is a discrete random variable B_j taking values in the set $\{b_j(1), b_j(2),..., b_j(n_j)\}$ with respective probabilities $p_j(1), p_j(2),..., p_j(n_j)$, where $0 \le b_j(1) < b_j(2) < \cdots < b_j(n_j) < \infty$. Let $\mathbf{B} = (B_1,..., B_a)$. A point $\mathbf{x}$ in the capacity state space Ω is defined as an a-tuple of values $\mathbf{x} = (b_1(v_1), b_2(v_2),..., b_a(v_a))$ where the index v_j takes on values from 1 to n_j. For notational convenience, the index v_j will also be used to designate the value $b_j(v_j)$ itself so that the state point $\mathbf{x}$ will also be denoted as $\mathbf{v} = (v_1,..., v_a)$.

Suppose that the capacities of the arcs are statistically independent. Therefore the probability $P(\mathbf{v})$ that state $\mathbf{v}$ occurs can be written as $P(\mathbf{v}) = \Pi_{j=1}^{a} p_j(v_j)$. For $\mathbf{v} \in \Omega$, let $L(\mathbf{v})$ denote the value of maximum s-t flow when the capacities are $\mathbf{v}$. Also, let $Z(C, \mathbf{v}) = \Sigma_{j \in C} b_j(v_j)$ denote the capacity of a (directed) s-t cut C. The cut C is minimum if $Z(C, \mathbf{v}) = L(\mathbf{v})$. Without loss of generality, we assume that every arc belongs to at least one s-t cut.

Suppose that demand d is placed at the sink t. Then $g(d) = P(L(\mathbf{B}) \ge d)$ is the network reliability. An arc j is called *critical* (with respect to the demand d) if it belongs to a minimum cut while $L(\mathbf{B}) < d$. We discuss the computation of:

- The network reliability.
- The probability $h(j)$ that arc j is critical. We refer to this probability as the *criticality index* of j.
- The probability $h(C)$ that a cut C is minimum. We refer to this probability as the *criticality index* of C.

A state $\mathbf{v} = (v_1,..., v_a)$ is called *operating* if flow of value d can be supplied from s to t when the capacities are $(b_1(v_1),..., b_a(v_a))$ and *failed* if no such *feasible* flow exists. If Ω^+ is the set of operating states, then $g(d) = P(\mathbf{B} \in \Omega^+)$. A set $R \subseteq \Omega$ is called operating or failed if all states in R are classified as operating or failed respectively.

Since the evaluations of $g(d)$ and $h(j)$ are *NP*-hard problems (Ball 1987, and Alexopoulos and Fishman 1991), no polynomial algorithm is known to exist for computing them. Exact methods for computing performance measures for flow networks include Doulliez and Jamoulle (1972), Evans (1976), Kulkarni and Adlakha (1985), Lee (1980), Rueger (1986), and Shogan (1982) while Monte Carlo methods are described in Alexopoulos (1993), Alexopoulos and Fishman (1991, 1993) and Fishman and Shaw (1989).

This note has the following two objectives: (1) It identifies and corrects errors in the algorithms by Doulliez and Jamoulle (1972) for computing the above probabilities; (2) It discusses the applicability of these procedures.

The methodology of Doulliez and Jamoulle (1972) is based on iteration and, in short, evaluates $g(d)$ as follows: At each iteration a subset of Ω is partitioned into non-overlapping operating, failed and undetermined sets. An undetermined set is one whose states cannot be classified at the current iteration as operating or failed. The existing undetermined sets are used as input to subsequent iterations. The method continues in the same fashion until all the undetermined sets have been considered. These sets are a-dimensional discrete rectangles in the sense that each such set, say R, has

limiting points $\boldsymbol{\alpha} \equiv \boldsymbol{\alpha}[R] = (\alpha_1,..., \alpha_a)$ and $\boldsymbol{\beta} \equiv \boldsymbol{\beta}[R] = (\beta_1,..., \beta_a)$ such that every integer vector $\mathbf{v}$ with $\alpha_j \le v_j \le \beta_j$ for all j belongs to R. This rectangle can then be denoted by $R = \{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$. Since the capacities are independent, the $P(\mathbf{B} \in R)$ can be written as

$$P(R) \equiv P(\mathbf{B} \in R) = \prod_{j=1}^{a} \sum_{\ell=\alpha_j}^{\beta_j} p_j(\ell).$$

Section 2 describes the modified algorithm for evaluating the network reliability $g(d)$ Section 3 contains the corrected algorithm for computing $h(j)$ and discusses the evaluation of $h(C)$. Section 4 contains conclusions and recommendations.

2. Computing the Network Reliability

Consider an undetermined rectangle $R \subseteq \Omega$ with lower and upper limiting points $\boldsymbol{\alpha} = (\alpha_1,..., \alpha_a)$ and $\boldsymbol{\beta} = (\beta_1,..., \beta_a)$ respectively. This set is partitioned by determining indices v_j^0 and v_j^* for each arc j such that the states $\mathbf{v} \in R$ for which $v_j^0 \le v_j \le \beta_j$, $j \in \mathcal{A}$ are operating and the states $\mathbf{v} \in R$ for which $v_j < v_j^*$ for some j are failed. These indices are obtained as follows: Create a fictitious demand node T, add the arc $e = (t, T)$ with capacity d and determine a maximum s-T flow $\mathbf{f}^0 = (f_1^0,..., f_a^0, f_e^0)$ with capacities $\boldsymbol{\beta}$ for the arcs in $\mathcal{A}$. If the value of this flow is less than d, then none of the states in R can satisfy the demand at node t making R a failed rectangle. Otherwise, for each arc $j \in \mathcal{A}$ define $v_j^0 = \min\{v : \alpha_j \le \ell \le \beta_j \text{ and } b_j(\ell) \ge f_j^0\}$. Obviously, all states $\mathbf{v}$ with $v_j^0 \le v_j \le \beta_j$, $j \in \mathcal{A}$ form an operating rectangle W.

We now show how the indices v_j^* can be obtained: Using the existing flows f_j^0, we compute a maximum s-t flow with value $L(\boldsymbol{\beta})$ and a minimum cut C. For each arc j with $v_j^0 = \alpha_j$ we set $v_j^* = \alpha_j$. Then for each arc $j = (k, l)$ with $v_j^0 > \alpha_j$ we identify a minimum s-t cut, say C', *containing* j ($C' = C$ for $j \in C$) and define v_j^* to be the smallest index $m \in \{\alpha_j, \alpha_j + 1,..., v_j^0\}$ such that $b_j(m) + \Sigma_{i \in C' - \{j\}} b_i(\beta_i) \ge d$. Clearly,

every state $\mathbf{v} \in R$ with $v_j < v_j^*$ is failed because the cut C' has capacity $Z(C', \mathbf{v}) < d$. For $j \notin C$, the cut C' is a minimum cut in the network resulting from G by adding two arcs (s, k) and (l, t) with infinite capacities and contains arc j because (1) the arcs in G have finite capacities and (2) j belongs to at least one s-t cut.

Remark 1: Doulliez and Jamoulle (1972, top of p. 55) proposed the following procedure for deriving the indices v_j^*: Given the feasible flow $\mathbf{f}^0$, for each arc $j = (k, l) \in \mathcal{A}$ let L_j denote the value of maximum flow that can be transmitted from node k to node l in addition to the existing flow without using this arc. If $L_j \geq f_j^0$, any capacity $\alpha_j \leq \ell \leq \beta_j$ for arc j satisfies the demand d when all other capacities are fixed at β_k, $k \neq j$ and then $v_j^* = \alpha_j$. Otherwise, arc j is in a minimum cut that blocks the value of maximum s-t flow below d and every state $\mathbf{v} \in R$ with $b_j(v_j) < f_j^0 - L_j$ is failed. Unfortunately, the last statement is not valid in general as the following example demonstrates. We should also mention that the statement remains false when L_j is the value of a maximum s-l flow or a maximum k-t flow.

The network in Figure 1 has $s = 1$, $t = 4$, demand $d = 6$, and arcs numbered as 1, 2,..., 5. Suppose that the capacity of each arc takes values from the set $\{1, 2, 3, 4, 5\}$. This assumption implies $b_j(\ell) = \ell$ for every $j \in \mathcal{A}$ and $\ell = 1,..., 5$. Consider the rectangle R with limiting points $\alpha = (1, 1, 1, 1, 1)$ and $\beta = (5, 3, 2, 3, 4)$. The s-t flow $\mathbf{f}^0 = (3, 3, 1, 2, 4)$ has value d and, clearly, $\mathbf{v}^0 = (3, 3, 1, 2, 4)$. Consider the arc $2 = (1, 3)$. The maximum flow that can still go from node 1 to node 3 without using arc 2 has value $L_2 = 1 < f_2^0$ (1 unit through arcs 1 and 3) and then $f_2^0 - L_2 = 2$. Now, reduce the capacity of arc 2 to $f_2 - L_2 - 1 = 1 \geq \alpha_2$ and leave the remaining capacities at β_k, $k \neq 2$. The flow $(5, 1, 2, 3, 3)$ has value $d = 6$ and the capacity state $(5, 1, 2, 3, 4) \in R$ is operating. The same erroneous statement reappears in Shogan (section V.C, step b). To the best of our knowledge, this error has not been pointed previously. □

Now for $j \in \mathcal{A}$ define

$$F'_j = \{\mathbf{v} : \mathbf{v} \in R \text{ and } v_j < v^*_j\} \ (= \emptyset \text{ if } v^*_j = \alpha_j)$$

and

$$U'_j = \{\mathbf{v} : \mathbf{v} \in R,\ v_j < v^0_j \text{ and } v_k \geq v^*_k \ \forall\, k \in \mathcal{A}\} \ (= \emptyset \text{ if } v^*_j = v^0_j).$$

Note that the sets F'_j are failed and that the states in $\cup_{j=1}^{a} U'_j$ cannot be classified as operating or failed with the existing flow. Since U_j overlap, we partition this union into the rectangles

$$\begin{aligned} U_j &= U'_j \setminus \cup_{k<j} (U_k \cap U'_j) \\ &= \{(v^0_1,\ldots, v^0_{j-1}, v^*_j, v^*_{j+1},\ldots, v^*_a); (\beta_1,\ldots, \beta_{j-1}, v^0_j - 1, \beta_{j+1},\ldots, \beta_a)\}. \end{aligned}$$

Similarly, we partition $\cup_{j=1}^{a} F'_j$ into the rectangles

$$\begin{aligned} F_j &= F'_j \setminus \cup_{k<j} (F_k \cap F'_j) \\ &= \{(v^*_1,\ldots, v^*_{j-1}, \alpha_j, \alpha_{j+1},\ldots, \alpha_a); (\beta_1,\ldots, \beta_{j-1}, v^*_j - 1, \beta_{j+1},\ldots, \beta_a)\}. \end{aligned}$$

Clearly, the rectangles W, U_j and F_j partition R.

Now, each non-empty undetermined rectangle U_j remains to be decomposed similarly to the original set R. Procedure RELIABILITY summarizes the decomposition method for calculating $g(d)$. The set Γ contains all the remaining undetermined rectangles. At termination the bounds $g_L(d)$ and $g_U(d)$ are equal to $g(d)$. Note that only the boundary points of each set must be stored. Efficient algorithms for maximum flow evaluations are described in Ahuja, Magnanti, and Orlin (1993).

Procedure RELIABILITY

Step 1 Start with the rectangle $R = \Omega$. Set $\Gamma = \emptyset$, $g_L(d) = 0$, and $g_U(d) = 1$.

Step 2 Let α and β denote the limiting points of R. If $L(\beta) < d$, set $\Gamma = \Gamma \backslash \{R\}$ and go to step 5. Otherwise, decompose R into an operating rectangle W, failed rectangles F_j, and undetermined rectangles U_j.

Step 3 Set $g_L(d) = g_L(d) + P(W)$.

Step 4 For $j \in \mathcal{A}$:

If $F_j \neq \emptyset$, set $g_U(d) = g_U(d) - P(F_j)$.

If $U_j \neq \emptyset$, set $\Gamma = \Gamma \cup \{U_j\}$.

Step 5 If $\Gamma \neq \emptyset$, choose a set from Γ, call it R, set $\Gamma = \Gamma \backslash \{R\}$, and go to step 2.

Step 6 End with $g(d) = g_L(d) = g_U(d)$.

The application of this procedure to the network in Doulliez and Jamoulle resulted in reliability 0.8824 which is larger, as expected, than their value of 0.8582. This evaluation required 96 rectangle decompositions and was verified by a Monte Carlo experiment.

Remark 2: The method of Doulliez and Jamoulle (1972) has been criticized for having to maintain large lists of undetermined sets; see for example Lee (1980) and Rueger (1986). Based on computational experience with state space partitioning methods for computing a variety of measures in several models (such as maximum flow, shortest path, and minimum spanning tree models), we find this claim unjustified. If procedure RELIABILITY is executed entirely, we recommend that Γ be maintained by means of a stack (or LIFO list). Using this strategy during the evaluation of unconditional measures, we never had to maintain more than 15-20 sets at any time for networks with up to 20 nodes. If high capacity levels k are associated with large probabilities $p_j(k)$ and the objective is a quick derivation of tight bounds, then we recommend that these collections be implemented as

heaps with nodes corresponding to rectangles and node weights equal to the negatives of the rectangle probabilities. Obviously, each procedure removes the set corresponding to the root of its respective heap.

In the special case where each component has only two states, an operating state and a failed state, an undetermined set can be stored by recording only the set of operating components and the set of failed components. □

3. Computing Criticality Indices

Our method for computing the criticality indices is based on the partition of the failed rectangles produced by procedure RELIABILITY. Indeed, consider a failed rectangle $F = \{\alpha, \beta\}$ generated in step 2 and, with capacities β, find a maximum s-t flow $\mathbf{f}$ and a minimum cut C. Then we use this flow to identify indices v_j^0, $j \in \mathcal{A}$ such that the states in the *classified* rectangle

$$D = \{\mathbf{v}: v_j^0 \le v_j \le \beta_j \forall\, j \in \mathcal{A}\}$$

have an identical set S of critical. This set consists of the arcs in C and the arcs $j \notin C$ with $f_j = b_j(\beta_j)$ for which a minimum s-t cut containing j (found by the method in Section 2) has capacity equal to $L(\beta)$. Then $P(D)$ is part of $h(j)$ only for $j \in S$. If $S = C$, we set $v_j^0 = \alpha_j$, $j \in S$; otherwise, we set $v_j^0 = \beta_j$, $j \in S$. Now for each arc $j \notin S$, we identify the index $k = \min\{\ell: \alpha_j \le \ell \le \beta_j$ and $b_j(\ell) \ge f_j\}$ and set $v_j^0 = k$ if j does not enter a minimum cut at level k or $v_j^0 = \min\{k+1, \beta_j\}$ otherwise. The partition of F ends with the determination of the *unclassified* rectangles

$$Q_j = \{(v_1^0, \ldots, v_{j-1}^0, \alpha_j, \alpha_{j+1}, \ldots, \alpha_a); (\beta_1, \ldots, \beta_{j-1}, v_j^0 - 1, \beta_{j+1}, \ldots, \beta_a)\}$$

for $j \in \mathcal{A}$ with $v_j^0 > \alpha_j$.

Procedure CRITICAL summarizes the preceding discussion. At termination, the bounds $h_L(j)$ and $h_U(j)$ are equal to $h(j)$. The set Δ contains the unclassified rectangles.

Procedure CRITICAL

Step 1 Start with the collection $\Delta = \{F_1,..., F_L\}$ of failed rectangles determined by procedure RELIABILITY. Set $h_L(j) = 0$ and $h_U(j) = 1 - g(d)$ for all $j \in \mathcal{A}$.

Step 2 If $\Delta = \emptyset$, go to step 5. Otherwise, remove a rectangle $R = \{\alpha, \beta\}$ from Δ and set $\Delta = \Delta \backslash \{R\}$. Decompose R into a classified rectangle D and unclassified rectangles Q_j. Let S be the set of critical arcs for the state β.

Step 3 For $j \in S$: Set $h_L(j) = h_L(j) + P(D)$.

For $j \notin S$: Set $h_U(j) = h_U(j) - P(D)$.

Step 4 For $j \in \mathcal{A}$: If $Q_j \neq \emptyset$, set $\Delta = \Delta \cup \{Q_j\}$.

Go to step 2.

Step 5 End.

Observe that procedure CRITICAL can be time consuming. In fact, it starts with Δ contains all failed rectangles determined by procedure RELIABILITY.

Remark 3: Doulliez and Jamoulle (1972, bottom of p. 55) considered the evaluation of $h(j)$, $j \in \mathcal{A}$ and proposed the following decomposition of a failed set F: Find a maximum flow $\mathbf{f}$ and a minimum cut C with capacities at levels β and set $v_j^0 = \alpha_j$ for $j \in C$ and $v_j^0 = \min\{\ell : \alpha_j \leq \ell \leq \beta_j \text{ and } b_j(\ell) \geq f_j\}$ for $j \notin C$. Then they claimed that only arcs in C are in minimum cuts for each state in the resulting set D. This statement is wrong as: (1) an arc $j \notin C$ may belong to an alternative minimum cut for the state β or (2) j may enter a minimum cut when its capacity is at level $v_j^0 < \beta_j$.

This error appears to affect the computation of the criticality index of a cut C as Doulliez and Jamoulle (1972, top of p. 56) state that this probability can be computed by summing up the probabilities of all sets D with C as the minimal cut. Unfortunately, the approach in the previous paragraph may fail to identify C as a minimum cut. Procedure CRITICAL does not seem to be applicable to the computation of the criticality indices of all cuts because the identification of *all* cuts with arcs in the set S is a hard problem. It appears that a separate decomposition procedure must be executed for computing the criticality index of each cut. Another alternative is the simultaneous estimation of the criticality indices of several cuts by using the method in Alexopoulos and Fishman (1993). □

4. Conclusions

The purpose of this note was to correct state space decomposition algorithms proposed by Doulliez and Jamoulle (1972) for the evaluation of performance characteristics of probabilistic transportation networks. Those algorithms are frequently referenced or used by researchers in the areas of power and communication systems and appear to be very effective for the computation of the network reliability when the demand is close to the largest possible maximum flow value.

We believe that extensive testing is required before the Jamoulle-Doulliez algorithms are disposed in favor of alternative approaches. Such testing should compare the performance of existing methods in a variety of networks including grid networks and dense networks of various sizes.

References

R.K. Ahuja, T.L. Magnanti, and J.B. Orlin. *Network Flows: Theory, Algorithms, and Applications.* Prentice Hall, 1993.

C. Alexopoulos. Computing criticality indices of arcs and the mean maximum flow value in networks with discrete random capacities, Technical Report J-93-01, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia, 1993.

C. Alexopoulos and G.S. Fishman. Characterizing stochastic flow networks using the Monte Carlo method. *Networks*, 21, 1991, 775-798.

C. Alexopoulos and G.S. Fishman. Sensitivity analysis in stochastic flow networks using the Monte Carlo method. *Networks*, 1993, forthcoming.

M.O. Ball. Computational complexity of network reliability analysis: An overview. *IEEE Transactions on Reliability*, 35, 1987, 230-239.

P. Doulliez and E. Jamoulle. Transportation networks with random arc capacities. *R.A.I.R.O.*, 3, 1972, 45-60.

J.R. Evans. Maximum flow in probabilistic graphs – the discrete case. *Networks*, 6, 1976, 161-183.

G.S. Fishman and T.Y. Shaw. Evaluating reliability of stochastic flow networks. *Probability in the Engineering and Informational Sciences*, 3, 1989, 493-509.

V.G. Kulkarni and V.G. Adlakha. Maximum flow in planar networks with exponential arc-capacities. *Communications in Statistics – Stochastic Models*, 1, 1985, 263-289.

S.H. Lee. Reliability of a flow network. *IEEE Transactions on Reliability*, 29, 1980, 24-26.

W.J. Rueger. Reliability analysis of networks with capacity constraints and failures at branches and nodes. *IEEE Transactions on Reliability*, 35, 1986, 523-528.

A.W. Shogan. Modular decomposition and reliability computation in stochastic transportation networks having cutnodes. *Networks*, 12, 1982, 255-275.

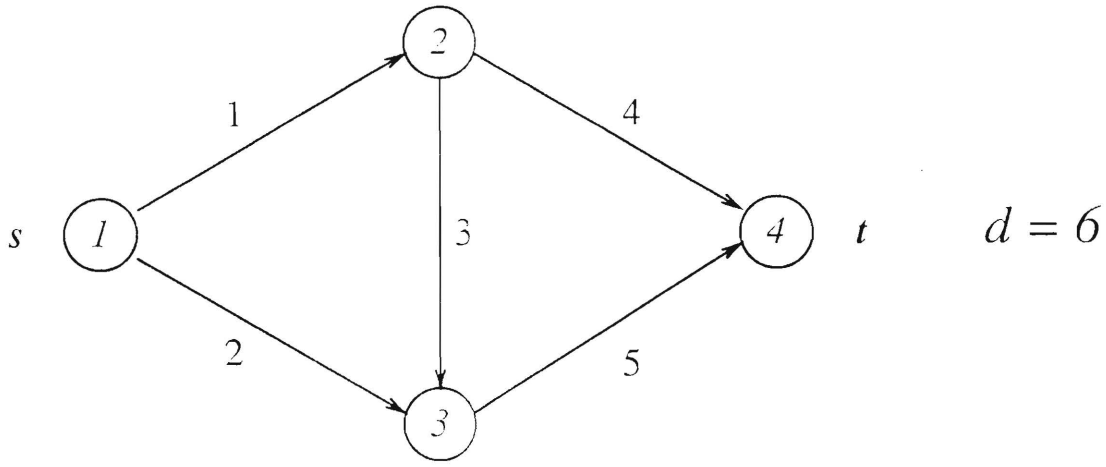

Figure 1

DISTRIBUTION – FREE CONFIDENCE INTERVALS FOR CONDITIONAL PROBABILITIES AND RATIOS OF EXPECTATIONS

CHRISTOS ALEXOPOULOS

To appear in *MANAGEMENT SCIENCE*

School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332-0205

This research was supported by the Air Force Office of Scientific Research under grant F49620-93-1-0043. Reproduction in whole or part is permitted for any purpose of the United States Government.

Many simulation experiments are concerned with the estimation of a ratio of two unknown means, the estimation of a conditional probability being an example. We propose confidence intervals for the case in which the ratio is estimated by using independent, identically distributed random pairs with bounded and ordered components. Emphasis is given to the case in which each component can be expressed as the product of a Bernoulli and a bounded random variable. The proposed intervals result from distribution-free bounds on error probabilities, are valid for every sample size, and their asymptotic width decreases at the same rate as that of confidence intervals based on the central limit theorem. We evaluate their performance by means of two experiments. The first considers the estimation of the probability that a path in a directed network is shortest while the second considers the estimation of the distribution of the inventory level in a stationary inventory system with periodic review. The experiments show that the intervals are conservative with superior coverage for small sample sizes (≤ 50).

(CONFIDENCE INTERVAL; SIMULATION; MONTE CARLO METHOD)

Introduction

Suppose that $\{(X_i, Y_i),\ i = 1,\ldots,n\}$ are independent and identically distributed (i.i.d.) random vectors, with generally dependent components, defined on a sample space Ω. Assume that $0 \le X_i \le Y_i \le m$, where m is a finite constant. Let $\mu_x = E(X_i)$, $\mu_y = E(Y_i)$ and suppose that $0 < \mu_x < \mu_y$. This paper develops confidence intervals for the ratio $\mu = \mu_x/\mu_y$. These confidence intervals are valid for any sample size n and their asymptotic width decreases at the same rate as for confidence intervals based on the central limit theorem. They are also conservative and their coverage is close to unity regardless of the nominal coverage requested by the user. The latter property limits their use in procedures whose objective is to control the coverage probability or satisfy some precision requirement.

Let $\overline{X}$ and $\overline{Y}$ denote the sample means of the X_i and Y_i respectively. Then

$$\bar{\mu} = \begin{cases} \overline{X}/\overline{Y} & \text{if } \overline{Y} > 0 \\ 0 & \text{if } \overline{Y} = 0 \end{cases} \tag{1}$$

is an estimator of μ.

There exist several cases, especially in Monte Carlo simulation, where the random variables (r.v.'s) X_i and Y_i can be expressed as

$$X_i = \phi_i \psi_i W_i \quad \text{and} \quad Y_i = \phi_i W_i; \quad 0 \le W_i \le m, \tag{2}$$

where ϕ_i and ψ_i are Bernoulli r.v.'s and W_i is a bounded r.v. Furthermore, ϕ_i, ψ_i and W_i are not necessarily independent. As an example, consider a finite sample space Ω with outcome probabilities $p(\omega)$, $\omega \in \Omega$ and the problem of estimating the conditional probability $\mu(p) = P_p(A|B)$, where A and B are two events with $P_p(B) > 0$. Let $\omega^{(1)},\ldots,\omega^{(n)}$ be n independent samples from Ω and define the Bernoulli variables

$$\phi_i = 1(\omega^{(i)} \in B), \quad \psi_i = 1(\omega^{(i)} \in A) \quad i = 1,\ldots,n,$$

where $1(\cdot)$ denotes the indicator function. Then $\mu(p)$ can be estimated by $\bar{\mu}$ in (1), where $X_i = \phi_i\psi_i = 1(\omega^{(i)} \in A \cap B)$ and $Y_i = \phi_i$. Now suppose that we want to use the samples from $p(\cdot)$ to estimate $\mu(q) = P_q(A|B) = E_q(\phi_i\psi_i)/E_q(\phi_i)$ for a different set of outcome probabilities $q(\omega)$ such that $p(\omega) > 0$ when $q(\omega) > 0$. Define $W_i = q(\omega^{(i)})/p(\omega^{(i)})$, $X_i = \phi_i\psi_i W_i$ and $Y_i = \phi_i W_i$, and note that the *weighting factors* W_i are used to ensure that $E_p(X_i) = E_q(\phi_i\psi_i)$, $E_p(Y_i) = E_q(\phi_i)$ and, therefore, $E_p(X_i)/E_p(Y_i) = \mu(q)$. If $\bar{Y} > 0$, then $\bar{\mu}(q,p) = \bar{X}/\bar{Y}$ is an estimate of $\mu(q)$ when the outcome probabilities are $q(\cdot)$. Note that W_i is bounded from above by $m = \sup\{q(\omega)/p(\omega)\colon \omega \in \Omega \text{ and } p(\omega) > 0\} < \infty$ since Ω is finite.

Problems of this nature arise in a variety of stochastic network settings. For example, consider a network whose arcs have random nonnegative lengths. We say that the network *functions* if there is a path from a source node v_s to a terminal node v_t whose length does not exceed a fixed value d. The probability that a given (v_s,v_t) path is shortest given that the network is in a functioning state measures the contribution of the path to the system performance. This model is revisited in Section 5.

The proposed confidence intervals can also be applied to the estimation of steady-state probabilities of regenerative processes (Heyman and Sobel 1982, pp. 179-193). Indeed, let $Z \equiv \{Z(t)\colon t \geq 0\}$ be a regenerative process with the following properties: (a) with probability one, its sample paths are right-continuous, have limits from the left, and make a finite number of jumps in each finite time interval; (b) its regeneration epochs are $0 = T_0 < T_1 < T_2 < \cdots$; and (c) the ith cycle length $T_{i+1} - T_i$ $(i = 0,1,\ldots)$ is a bounded nonarithmetic r.v. Let S denote the state space of Z and let B denote the class of Borel sets of S. Then for $A \in B$, the limiting probability that the process is in the set A is

$$\mu(A) = \lim_{t\to\infty} P(Z(t) \in A) = E\left[\int_0^{T_1} 1(Z(s) \in A)ds\right]/E(T_1)$$

and can be estimated by simulating the process over n cycles. Let X_i be the total time the process spends in A during cycle i and let Y_i be the length of this cycle. Then $\mu(A)$ can be estimated by $\overline{X}/\overline{Y}$. If T_1 is arithmetic, then the above equation holds when t is a multiple of the span.

As an example, consider an inventory system operating under a stationary (s,S) policy (Heyman and Sobel 1982, pp. 134-135). Assume that the demands during periods are i.i.d. discrete r.v.'s taking on positive values and orders are filled immediately. Let I_i be the level of inventory on hand plus on order in period i immediately after the ordering decision and assume that $I_0 = S$. Then $\{I_i: i \geq 0\}$ is a (discrete-time) regenerative process with demarcating state S. Unfortunately, the restriction that the random variables Y_i be bounded prohibits the application of the proposed confidence intervals to the majority of queueing simulations that use the regenerative approach for estimating steady-state measures.

This paper considers two cases: (a) the *basic* case in which $X_i = \phi_i \psi_i W_i$, $Y_i = \phi_i W_i$ and W_i is a bounded r.v.; and (b) the *general* case in which X_i and Y_i follow an arbitrary joint distribution. Section 2 describes existing approaches for computing confidence intervals. Section 3 proposes statistical inequalities and uses them for developing the confidence intervals. Algorithms for computing these intervals are listed. It is shown that the confidence interval for the basic case is a subset of the confidence interval for the general case. Section 4 extends the results in Section 3 for the case in which $X_i = \phi_i \psi_i V_i$, $Y_i = \phi_i W_i$ and the ratio V_i/W_i is bounded. Section 5 evaluates the performance of the intervals by means of a stochastic shortest path model and an (s,S) inventory system with periodic review. Finally, Section 6 contains conclusions and recommendations.

2. Existing Confidence Intervals

The traditional approach for deriving confidence intervals for ratios uses the central limit theorem (CLT). The approach considers the i.i.d. r.v.'s

$$D_i = X_i - \mu Y_i \qquad i = 1, \ldots, n$$

with sample mean $\overline{D}$. Note that $E(D_i) = 0$ and $\sigma^2 = \text{var}(D_i) = \sigma_x^2 - 2\mu\sigma_{xy} + \mu^2\sigma_y^2$, where $\sigma_x^2 = \text{var}(X_i)$, $\sigma_y^2 = \text{var}(Y_i)$ and $\sigma_{xy} = \text{cov}(X_i, Y_i)$. Let $s^2(X)$ and $s^2(Y)$ denote the sample variances of X_i and Y_i respectively, and let $s(X,Y)$ denote the sample covariance of X_i and Y_i. By the CLT, $\overline{D}/[\text{var}(\overline{D})]^{1/2}$ asymptotically has the standard normal distribution and for large n

$$P\left[\sqrt{n}|\overline{D}|/\sigma \le z_{1-\alpha/2}\right] \simeq 1 - \alpha, \tag{3}$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution.

The classical, and most commonly used approach, estimates σ^2 by $\hat{s}^2(D) = s^2(X) - 2\overline{\mu}s(X,Y) + \overline{\mu}^2 s^2(Y)$ to produce the (approximate) $1-\alpha$ confidence interval

$$\overline{\mu} \pm z_{1-\alpha/2}\, \hat{s}(D)/(\overline{Y}\sqrt{n}).$$

The method in Fieller (1954) replaces σ^2 by $s^2(X) - 2\mu s(X,Y) + \mu^2 s^2(Y)$ and solves the resulting inequality in (3) to yield the confidence interval

$$\frac{\overline{X}\cdot\overline{Y} - cs(X,Y)}{\overline{Y}^2 - cs^2(Y)} \pm \frac{A^{1/2}}{\overline{Y}^2 - cs^2(Y)},$$

where

$$A = [\overline{X}\cdot\overline{Y} - cs(X,Y)]^2 - [(\overline{X})^2 - cs^2(X)][(\overline{Y})^2 - cs^2(Y)] \quad \text{and} \quad c = z_{1-\alpha/2}^2/n.$$

For a small sample size n, Iglehart (1975) showed that the approximate confidence interval

$$\bar{\mu}_J \pm z_{1-\alpha/2}\sqrt{s_J^2/n},$$

where

$$\bar{\mu}_J = (1/n)\sum_{i=1}^{n}\theta_i; \quad \theta_i = n(\bar{X}/\bar{Y})-(n-1)(\sum_{j\neq i} X_j)/(\sum_{j\neq i} Y_j)$$

is the jackknife estimator of μ and

$$s_J^2 = \sum_{i=1}^{n}(\theta_i-\bar{\mu}_J)^2/(n-1),$$

often provides better coverage than the first two confidence intervals. However, its evaluation requires substantial bookkeeping in addition to $O(n^2)$ operations, making its use costly for large sample sizes.

As a check on the normality assumption, we can compute an estimate of the skewness

$$\rho(X,Y) = E(D_i^3)/[\mathrm{var}(D_i)]^{3/2}.$$

When the absolute value of this measure is large, the convergence of $\bar{D}/[\mathrm{var}(\bar{D})]^{1/2}$ to the standard normal distribution is usually slow. When X_i and Y_i are ordered Bernoulli variables, Fishman (1987) shows that the Fieller confidence interval covers the true mean μ with probability that is often less than the desired $1-\alpha$, even for large n.

To avoid the dependence on the asymptotic normality assumption, we now turn to the derivation of nonnormal confidence intervals. Chebyshev's inequality

$$P[|\bar{X}-\mu\bar{Y}| \geq \epsilon] \leq \frac{\sigma^2}{n\epsilon^2} = \frac{\sigma_x^2-2\mu\sigma_{xy}+\mu^2\sigma_y^2}{n\epsilon^2}$$

provides a bound on both probabilities which, unfortunately, contains the unknowns σ_x^2, σ_y^2 and σ_{xy}. In addition, the bounds in Section 3 are typically better than the Chebyshev bound.

A special case arises when X_i and Y_i are Bernoulli r.v.'s. Let $S_x = \Sigma_{i=1}^n X_i$, $S_y = \Sigma_{i=1}^n Y_i$ and assume that $S_y > 0$. Then, conditional on $S_y = k$, S_x has the binomial(k,μ) distribution with cumulative distribution function (c.d.f.) $H(j;k,\mu) = P(S_x \le j \mid S_y = k)$ for $0 \le j \le k$. Let $\mu_L(j,k,\alpha_L)$ and $\mu_U(j,k,\alpha_U)$ denote the (unique) solutions to the equations

$$1-H(j-1;k,r) = \alpha_L, \quad 0 < r < 1$$

and (4)

$$H(j;k,r) = \alpha_U, \quad 0 < r < 1$$

respectively, where α_L, $\alpha_U \in (0,1)$ are such that $\alpha = \alpha_L+\alpha_U \in (0,1)$. Then, conditionally on $S_y = k$, $(\mu_L(S_x,k,\alpha_L),\mu_U(S_x,k,\alpha_U))$ is a confidence interval for μ with coverage at least equal to $1-\alpha$ and hence $(\mu_L(S_x,S_y,\alpha_L), \mu_U(S_x,S_y,\alpha_U))$ is an unconditional confidence interval for μ with coverage at least equal to $1-\alpha$, where $\mu_L(S_x,S_y,\alpha_L) \equiv 0$ and $\mu_U(S_x,S_y,\alpha_U) \equiv 1$ when $S_y = 0$. Equation system (4) can be solved by noting that (Abramowitz and Stegun 1964, p. 945)

$$1-H(j-1;k,r) = \int_0^r \frac{k!}{(j-1)!(k-j)!} z^{j-1}(1-z)^{k-j}dz.$$

Therefore $\mu_L(S_x,S_y,\alpha_L)$ calls for the evaluation of the inverse Beta distribution with parameters S_x and S_y-S_x+1 and $\mu_U(S_x,S_y,\alpha_U)$ for the evaluation of the inverse Beta distribution with parameters S_x+1 and S_y-S_x. Although these evaluations can be performed by using subroutines in IMSL (1982) or Press *et al.* (1989), experience with these programs indicates numerical difficulties when S_y is large. The α_L and α_U that produce a

$1-\alpha$ confidence interval with minimal width can be chosen heuristically with a grid search in the interval $(0,\alpha]$.

3. Proposed Confidence Intervals

The method employed to derive the proposed confidence intervals is based on upper bounds on the error probabilities $P[\overline{X}-\mu\overline{Y} \geq \epsilon]$ and $P[-\overline{X}+\mu\overline{Y} \geq \epsilon]$ for $\epsilon > 0$ which are functions of the first moments of X_i and Y_i only. The use of bounds for the derivation of confidence intervals is not new; see for example pp. 68-73 of Shiryayev (1984). To eliminate unnecessary notation, we assume that $m = 1$. If $m \neq 1$, we derive the confidence intervals by using the random variables $X_i^* = X_i/m$ and $Y_i^* = Y_i/m$ with $E(Y_i^*) = \mu_y/m$. Otherwise, the bounds contain m and the resulting confidence intervals are not affected if m is the right endpoint of the c.d.f. of Y_i. When m is an arbitrarily large bound, computational experience suggests that for $n \geq 50$ the interval for the basic case widens proportionally to $\sqrt{m}$ while the interval for the general case widens proportionally to m. These observations should make us skeptical about using an upper bound m that is considerably greater than the right endpoint of the c.d.f. of Y_i.

The approach used to derive the bounds has apparently been used first by S.N. Bernstein and leads to considerably tighter bounds than Chebyshev's inequality, as pointed out by Hoeffding (1963). We develop the confidence intervals in three steps. The first step develops the upper bounds. The second step proposes the confidence intervals. In particular, the interval for the basic case is parametric in $t \in [\mu_y,1]$, widens as t increases, and yields the interval for the general case when $t = 1$. The third step then uses Bonferroni's principle to propose a confidence interval by replacing t by a random upper confidence limit on μ_y. The Appendix contains the proofs of the theorems.

Our approach for deriving bounds on the error probabilities is based on the following simple observation. The probability $P[\Sigma_{i=1}^n(X_i-\mu Y_i-\epsilon) \geq 0]$ is the expected value of the indicator variable $1\{\Sigma_{i=1}^n(X_i-\mu Y_i-\epsilon) \geq 0\}$. Since this variable does not exceed

$\exp\{h\,\Sigma_{i=1}^{n}(X_i-\mu Y_i-\epsilon)\}$ for all $h \geq 0$, it follows that

$$P\,[\overline{X}-\mu\overline{Y} \geq \epsilon] \leq E\left[\exp\{h\,\Sigma_{i=1}^{n}(X_i-\mu Y_i-\epsilon)\}\right] = e^{-nh\epsilon}\left[E\,e^{h(X_1-\mu Y_1)}\right]^n, \tag{5}$$

where the last equality follows from the fact that (X_i, Y_i) are i.i.d. pairs. Similarly,

$$P\,[-\overline{X}+\mu\overline{Y} \geq \epsilon] \leq e^{-nh\epsilon}\left[E\,e^{h(-X_1+\mu Y_1)}\right]^n.$$

Theorem 1 below proposes upper bounds on $E\,e^{h(X_1-\mu Y_1)}$ and $E\,e^{h(-X_1+\mu Y_1)}$ and therefore upper bounds on $P\,[\overline{X}-\mu\overline{Y} \geq \epsilon]$ and $P\,[-\overline{X}+\mu\overline{Y} \geq \epsilon]$. Part (ii) of the theorem follows from Theorem 1 in Hoeffding when we write $P\,[\overline{X}-\mu\overline{Y} \geq \epsilon] = P\,[\overline{X}+\mu(1-\overline{Y})-\mu \geq \epsilon]$ and $P\,[-\overline{X}+\mu\overline{Y} \geq \epsilon] = P\,[\overline{Y}-\overline{X}+(1-\mu)(1-\overline{Y})-(1-\mu) \geq \epsilon]$. Part (iii) of Theorem 1 below extends Theorem 5 (i)-(ii) in Fishman (1989) where X_i and Y_i are Bernoulli variables.

Theorem 1. (*i*) *Define the function*

$$F(t,h,r,\epsilon) = e^{-h\epsilon}\left\{1+t\left[-1+\,r\,e^{(1-r)h}+(1-r)e^{-rh}\right]\right\}$$
$$\textit{for } 0 < t \leq 1,\ h \geq 0,\ 0 \leq r \leq 1, \textit{ and } 0 \leq \epsilon \leq 1-r.$$

For each $(t,r,\epsilon) \in S = \{(t,r,\epsilon)\colon 0 < t \leq 1,\ 0 < r < 1,\ 0 < \epsilon < 1-r\}$, *F is strictly convex with respect to h and has a unique positive and finite minimum. Furthermore, there exists a unique function*

$$h^*(t,r,\epsilon)\colon S \longrightarrow (0,\infty)$$

which is smooth (continuously differentiable) in the interior of S, $\mathrm{int}(S)$, *satisfies*

$$\frac{\partial F}{\partial h}(t,h^*(t,r,\epsilon),r,\epsilon) = 0 \quad \textit{for all } (t,r,\epsilon) \in \mathrm{int}(S),$$

and is given analytically for $t = 1$ *by*

$$h^*(t,r,\epsilon) = \log\left[\frac{(1-r)(r+\epsilon)}{r(1-r-\epsilon)}\right].$$

(ii) Suppose that the random vectors (X_i, Y_i), $i = 1,\ldots,n$ *are i.i.d. and* $0 \le X_i \le Y_i \le 1$. *Let* $\mu_x = E(X_i)$, $\mu_y = E(Y_i)$, *and* $\mu = \mu_x/\mu_y$ *and assume that* $0 < \mu < 1$ *and* $0 < \epsilon < 1-\mu$. *Then*

$$P[\overline{X}-\mu\overline{Y} \ge \epsilon] \le [F(1,h^*(1,\mu,\epsilon),\mu,\epsilon)]^n = \exp\{nG(\mu,\epsilon)\}, \tag{6}$$

where

$$G(r,\epsilon) = (r+\epsilon)\log[r/(r+\epsilon)] + (1-r-\epsilon)\log[(1-r)/(1-r-\epsilon)]. \tag{7}$$

Furthermore,

$$P[-\overline{X}+\mu\overline{Y} \ge \epsilon] \le [F(1,h^*(1,1-\mu,\epsilon),1-\mu,\epsilon)]^n = \exp\{nG(1-\mu,\epsilon)\}. \tag{8}$$

(iii) If in addition X_i *and* Y_i *are expressed as* $X_i = \phi_i\psi_i W_i$ *and* $Y_i = \phi_i W_i$, *where* ϕ_i *and* ψ_i *are Bernoulli r.v.'s, then*

$$E\, e^{h[X_1-\mu Y_1]} \le 1-\mu_y+\mu_y\left[\mu\, e^{(1-\mu)h} + (1-\mu)\, e^{-\mu h}\right]$$

and

$$P[\overline{X}-\mu\overline{Y} \ge \epsilon] \le [F(\mu_y,h^*(\mu_y,\mu,\epsilon),\mu,\epsilon)]^n. \tag{9}$$

Furthermore,

$$P[-\overline{X}+\mu\overline{Y} \ge \epsilon] \le [F(\mu_y,h^*(\mu_y,1-\mu,\epsilon),1-\mu,\epsilon)]^n. \tag{10}$$

□

Direct substitution of $h^*(1,\mu,\epsilon)$ and $h^*(1,1-\mu,\epsilon)$ into the r.h.s. of (9) and (10) shows that these bounds yield the r.h.s. of (6) and (8) respectively when μ_y is replaced by unity.

Remark 1. The bounds in parts (ii) and (iii) of Theorem 1 are the best that can be obtained from inequality (5) under the assumption $0 \leq X_i \leq Y_i \leq 1$. Indeed, if $Y_i = 1$ with probability one and X_i has the Bernoulli distribution $P(X_i = 1) = \mu$ and $P(X_i = 0) = 1-\mu$, then for $\epsilon = 1-\mu$ we have $P[\overline{X}-\mu\overline{Y} \geq \epsilon] = P[\Sigma_{i=1}^{n} X_i = n] = \mu^n$ while $\lim_{\epsilon \to 1-\mu} \exp\{G(\mu,\epsilon)\} = \mu$. Also, for fixed ϵ, these bounds typically approach zero at an exponential rate as n increases (see Hoeffding 1963, p. 14).

Theorem 2 considers the case in Theorem 1 (iii) and shows how the inequalities (9) and (10) can be used to derive a $1-\alpha$ confidence interval for μ. Corollary 1 then uses inequalities (6) and (8) to obtain a confidence interval for μ. Its proof results from that of Theorem 2 when t and μ_y are replaced by unity.

Theorem 2. (*i*) *For each* $0 < t < 1$, $0 < \beta < 1$, *and* $0 < u \leq v \leq 1$ *the system*

$$\begin{aligned} F(t,h^*(t,r,\epsilon),r,\epsilon) &= \beta^{1/n} \\ \epsilon &= u-vr \\ 0 &< r < u/v \end{aligned} \tag{11}$$

has a unique solution $0 < r'_\beta(t,u,v) < \min\{u/v, \beta^{1/n}/t\}$. *Furthermore, there are unique smooth functions*

$$\begin{aligned} \overline{h}(t,r)&: R \longrightarrow (0,\infty) \\ \overline{\epsilon}(t,r)&: R \longrightarrow (0,1) \end{aligned}$$

defined on $R = \{(t,r): 0 < t < 1, 0 < r < 1, tr < \beta^{1/n}\}$ *which satisfy*

$$\begin{aligned} F(t,\overline{h}(t,r),r,\overline{\epsilon}(t,r)) &= \beta^{1/n} \\ \frac{\partial F}{\partial h}(t,\overline{h}(t,r),r,\overline{\epsilon}(t,r)) &= 0 \end{aligned}$$

$$\bar{h}(t,r) = h^{*}(t,r,\bar{\epsilon}(t,r)) \qquad \text{for all } (t,r) \in R$$
$$\bar{\epsilon}(t,r) < 1\text{-}r.$$

For each t the curve $\bar{\epsilon}(t,r)$ is strongly quasiconcave in r.

(*ii*) *If $t = 1$, then for all $0 < \beta < 1$ and $0 < u \leq v \leq 1$ the system*

$$G(r,\epsilon) = (\log\beta)/n$$
$$\epsilon = u\text{-}vr \tag{12}$$
$$0 < r < u/v$$

has a unique solution $0 < r'_{\beta}(1,u,v) < \min\{u/v, \beta^{1/n}\}$. In addition, there is a unique, strictly concave and smooth function

$$\hat{\epsilon}(r)\colon (0,\beta^{1/n}) \longrightarrow (0,1)$$

which satisfies

$$G(r,\hat{\epsilon}(r)) = (\log\beta)/n \tag{13}$$
$$\hat{\epsilon}(r) < 1\text{-}r \qquad \text{for all } r.$$

(*iii*) *Let α, β, $\gamma \in (0,1)$ such that $\alpha = \beta+\gamma$. Assume that X_i and Y_i are defined as in Theorem* 1 (*iii*). *Define*

$$L(t,\beta) = \begin{cases} r'_{\beta}(t,\bar{X},\bar{Y}) & \text{if } \bar{X} > 0 \\ 0 & \text{if } \bar{X} = 0 \end{cases} \tag{14}$$

and

$$U(t,\gamma) = \begin{cases} 1\text{-}r'_{\gamma}(t,\bar{Y}\text{-}\bar{X},\bar{Y}) & \text{if } \bar{X} < \bar{Y} \\ 1 & \text{if } \bar{X} = \bar{Y}. \end{cases} \tag{15}$$

Then

$L(t,\beta)$ *is nonincreasing in t w.p.* 1,

$U(t,\gamma)$ *is nondecreasing in* t *w.p.* 1

and

$$P\,[L(t,\beta) < \mu < U(t,\gamma)] > 1-\alpha \quad \textit{for all } t \in [\mu_y,1]. \tag{16}$$

□

Corollary 1. *Let* α, β, $\gamma \in (0,1)$ *such that* $\alpha = \beta+\gamma$. *Assume that the random vectors* (X_i, Y_i) *are defined as in Theorem* 1 (*ii*). *Let* $L(\beta) \equiv L(1,\beta)$ *and* $U(\gamma) \equiv U(1,\gamma)$, *where* $L(1,\beta)$ *and* $U(1,\gamma)$ *are defined by* (14) *and* (15), *respectively, for* $t = 1$. *Then*

$$P\,[L(\beta) < \mu < U(\gamma)] > 1-\alpha.$$

□

Theorem 2 (iii) implies that for fixed β and γ the narrowest confidence interval is computed when t is equal to the unknown mean μ_y. Although we can derive an approximate confidence interval for μ by replacing μ_y by its estimate $\overline{Y}$, the algorithmic derivation of this interval makes it difficult to study the distortion in the confidence level which is induced by this substitution.

An alternative approach avoids this error of approximation. Suppose that $M_y(\delta)$, $0 < \delta < 1$ is a r.v. such that $M_y(\delta) \leq 1$ and $P\,[\mu_y \leq M_y(\delta)] > 1-\delta$. Then, by the Bonferroni principle and the fact that $L(t,\beta)$ (respectively, $U(t,\gamma)$) is nonincreasing (respectively, nondecreasing) in t, if follows immediately that the interval $(L(M_y(\delta),\beta), U(M_y(\delta),\gamma))$ covers μ with probability greater than $1-\alpha-\delta$.

An upper confidence limit on μ_y can be computed by using the following bound from Hoeffding (1963)

$$P\,[-\overline{Y}+\mu_y \geq \epsilon] \leq \exp\{nG(1-\mu_y,\epsilon)\},$$

where G is defined by (7).

Theorem 3 below uses these ideas to propose a confidence interval for μ. Inequality

(18) is from Theorem 1 in Fishman (1991). The remainder of the proof is obvious.

Theorem 3. *Let* $\alpha, \beta, \gamma, \delta \in (0,1)$ *so that* $\alpha = \beta+\gamma+\delta$. *Let* z_0 *denote the solution to*

$$G(1-z,-\overline{Y}+z) = (\log\delta)/n, \quad \overline{Y} < z < 1$$

when $\overline{Y} < 1$ *and let*

$$M_y(\delta) = \begin{cases} z_0 & \text{if } \overline{Y} < 1 \\ 1 & \text{if } \overline{Y} = 1 . \end{cases} \tag{17}$$

Then

$$P[\mu_y < M_y(\delta)] > 1-\delta \tag{18}$$

and

$$P[L(M_y(\delta),\beta) < \mu < U(M_y(\delta),\gamma)] > 1-\alpha,$$

where $L(t,\beta)$ *and* $U(t,\gamma)$ *are defined in* (14) *and* (15).

□

Algorithm A describes the computation of $M_y(\delta)$. If $\overline{Y} < 1$, then z_0 is computed by using the bisection method in the interval $(\overline{Y},1)$. This method can also be used for computing $r'_\beta(M_y(\delta), \overline{X},\overline{Y})$ and $r'_\gamma(M_y(\delta),\overline{Y}-\overline{X},\overline{Y})$. Indeed in the proof of Theorem 2 it is shown that, for fixed t, β and N, $F(t,h^*(t,r,\overline{X}-r\overline{Y}),r,\overline{X}-r\overline{Y})$ is increasing in $r \in (0,\overline{X}/\overline{Y})$. Similarly, it can be shown that $F(t,h^*(t,r,\overline{Y}-\overline{X}-r\overline{Y}),r,\overline{Y}-\overline{X}-r\overline{Y})$ is increasing in $r \in (0,1-\overline{X}/\overline{Y})$.

Algorithm B describes the computation of $L(M_y(\delta),\beta)$. Steps 2 b-c use the bisection method for computing $h^*(t,r,\epsilon)$. Bounds on $h^*(t,r,\epsilon)$ can be computed by noting that the inequalities

$$\frac{\partial F}{\partial h} \le e^{-h\epsilon}\left[-\epsilon(1-t)+tr(1-r-\epsilon)e^{(1-r)h}\right]$$

and

$$\frac{\partial F}{\partial h} \geq e^{-h\epsilon}\left[-\epsilon(1-t)+tr(1-r-\epsilon)e^{(1-r)h}-t(1-r)(r+\epsilon)\right]$$

imply

$$\frac{1}{1-r}\log\left[\frac{(1-t)\epsilon}{tr(1-r-\epsilon)}\right] = h_L^* \leq h^*(t,r,\epsilon) \leq h_U^* = \frac{1}{1-r}\log\left[1+\frac{\epsilon}{tr(1-r-\epsilon)}\right],$$

where h_L^* (h_U^*) is computed by equating the upper (lower) bound on $\frac{\partial F}{\partial h}$ to zero and solving for h. The upper confidence limit $U(M_y(\delta),\gamma)$ can be computed if we replace β by γ and set:

1. $r \leftarrow 1$ if $\overline{X} = \overline{Y}$ in the initialization step.
2. $r_L \leftarrow 0$ and $r_U \leftarrow 1-\overline{X}/\overline{Y}$ in step 1.
3. $\epsilon \leftarrow (\overline{Y}-\overline{X})-r\overline{Y}$ in step 2a.
4. $U(M_y(\delta),\gamma) \leftarrow 1-r$ in step 3.

FORTRAN 77 codes for both algorithms are available from the author upon request.

Algorithm A

Purpose: To compute $M_y(\delta)$ defined by equation (17)

Input: Sample size n; sample mean $\overline{Y}$; $0 < \delta < 1$; and error tolerance ξ_1

Output: $M_y(\delta)$

Method:

If $\overline{Y} = 1$: set $z \leftarrow 1$; go to step 3

Define: $H(z) = G(1-z,-\overline{Y}+z) - (\log\delta)/n$

1. $z_L \leftarrow \overline{Y}$, $z_U \leftarrow 1$
2. **repeat**

   begin

   a. $z \leftarrow (z_L+z_U)/2$;

   b. Compute $H(z_L)$ and $H(z)$;

   if $H(z_L)H(z) < 0$ **then** $z_U \leftarrow z$ **else** $z_L \leftarrow z$

   end

until $|H(z)| \leq \xi_1$

3. Deliver $M_y(\delta) \longleftarrow z$

Algorithm B

Purpose: To compute $L(M_y(\delta),\beta)$ defined by equation (14) for $t = M_y(\delta)$

Input: Sample size n; sample means $\overline{X}$ and $\overline{Y}$; $t = M_y(\delta)$ from Algorithm A; $0 < \beta < 1$; and error tolerances ξ_2 and ξ_3

Output: $L(M_y(\delta),\beta)$

Method:

If $\overline{X} = 0$: set $r \longleftarrow 0$; go to step 3

Define:

$$F = e^{-h\epsilon}\left\{1+t\left[-1+ r\, e^{(1-r)h}+(1-r)e^{-rh}\right]\right\}$$

$$F^* = \min_{h\geq 0} F$$

$$\partial F/\partial h = -\epsilon F + e^{-h(r+\epsilon)}tr(1-r)(e^h-1)$$

1. $r_L \longleftarrow 0$, $r_U \longleftarrow \overline{X}/\overline{Y}$

2. **repeat**

begin

a. $r \longleftarrow (r_L+r_U)/2$, $\epsilon \longleftarrow \overline{X}-r\overline{Y}$

b. Compute bounds on $h^*(t,r,\epsilon)$:

$$h_L^* \longleftarrow \frac{1}{1-r}\log\left[\frac{(1-t)\epsilon}{tr(1-r-\epsilon)}\right]$$

$$h_U^* \longleftarrow \frac{1}{1-r}\log\left[1+\frac{\epsilon}{tr(1-r-\epsilon)}\right]$$

c. **repeat**

begin

$h \longleftarrow (h_L^*+h_U^*)/2$;

Compute $\partial F/\partial h$;

if $\partial F/\partial h < 0$ **then** $h_L^* \longleftarrow h$ **else** $h_U^* \longleftarrow h$

end

until $|\partial F/\partial h| \leq \xi_3$

d. Compute F^*

if $F^* \leq \beta^{1/n}$ **then** $r_L \longleftarrow r$ **else** $r_U \longleftarrow r$

end

until $\beta^{1/n} - \xi_2 \leq F^* \leq \beta^{1/n}$

3. Deliver $L(M_y(\delta),\beta) \longleftarrow r$

The algorithmic derivation of the proposed intervals makes the determination of their asymptotic width a hard problem. Theorem 2 (iii) implies $L(\beta) \leq L_1(t,\beta) \leq U_1(t,\gamma) \leq U(\gamma)$ for all $t \in [\mu_y,1]$. Theorem 4 below proposes an upper bound on the width of the interval $(L(\beta),L(\gamma))$. Since $\overline{Y}$ converges to μ_y a.s. as $n \to \infty$, it follows that the width of our intervals is $O_P(n^{-1/2})$, where $\{O_P(n^{-1/2}),\ n = 1,2,...\}$ is a sequence of r.v.'s such that $O_P(n^{-1/2})/n^{-1/2}$ converges a.s. to a constant as $n \to \infty$.

Theorem 4. *Suppose* $0 < \overline{X} < \overline{Y}$. *Then*

$$U(\gamma) - L(\beta) \leq \left\{[(-\log\beta)/2]^{1/2} + [(-\log\gamma)/2]^{1/2}\right\}/(\overline{Y}\sqrt{n}). \tag{19}$$

□

The assignments of values to β, γ and δ which minimize the width of the confidence intervals in Theorem 3 and Corollary 1 seem to be hard problems. Computational experience has shown that for the first interval the choice $\delta = \overline{Y}\alpha/(\overline{\mu}+\overline{Y})$, $\beta = \overline{\mu}(\alpha-\delta)$ and $\gamma = \alpha-\delta-\beta$ is often preferable to the alternative $\delta = \alpha/2$ and $\beta = \gamma = \alpha/4$. Similarly, the assignment $\beta = \overline{\mu}\alpha$ and $\gamma = \overline{\mu}(1-\alpha)$ frequently yields a narrower interval $(L(\beta),L(\gamma))$ than the alternative $\beta = \gamma = \alpha/2$.

Assuming $\beta = \gamma = \alpha/2$, the ratio of the width in (19) to that of the classical confidence interval in Section 2 is bounded from above by

$$\frac{[-\log(\alpha/2)/2]^{1/2}}{z_{1-\alpha/2}\hat{s}(D)},$$

where $\hat{s}^2(D)$ is the estimator of $\mathrm{var}(D_i)$ defined in Section 2. The factor $[-\log(\alpha/2)/2]^{1/2}/z_{1-\alpha/2}$ decreases as $\alpha \downarrow 0$ and equals 0.74 for $\alpha = 0.10$, 0.69 for $\alpha = 0.05$, and 0.63 for $\alpha = 0.01$.

4. Unequal Weighting Factors

Suppose now that the variables X_i and Y_i are defined as in (2), but have unequal weighting factors $0 < V_i \leq 1$ and $0 < W_i \leq 1$, respectively, with bounded ratio. That is, suppose that

$$X_i = \phi_i \psi_i V_i, \quad Y_i = \phi_i W_i$$

where $i = 1,\ldots,n$

$$0 < c_1 \leq V_i/W_i \leq c_2 < \infty.$$

Let

$$Q_i = \phi_i \psi_i W_i, \quad R_i = \phi_i V_i$$

and note that for $\lambda = E(Q_i)/E(Y_i)$ and $\theta = E(X_i)/E(R_i)$,

$$\mu = \frac{E(X_i)}{E(Y_i)} = \frac{E(\phi_i \psi_i V_i)}{E(\phi_i W_i)} \geq c_1 \frac{E(\phi_i \psi_i W_i)}{E(\phi_i W_i)} = c_1 \lambda$$

and

$$\mu = \frac{E(X_i)}{E(Y_i)} \leq c_2 \frac{E(\phi_i \psi_i V_i)}{E(\phi_i V_i)} = c_2 \theta.$$

Suppose that $\Lambda(\beta)$ is a $1-\beta$ lower confidence limit for λ calculated by using the pairs (Q_i, Y_i) so that

$$P[\Lambda(\beta) \geq \lambda] \leq \beta$$

and that $\Theta(\gamma)$ is a $1-\gamma$ upper confidence limit for θ computed by using the pairs (X_i,R_i) so that

$$P[\Theta(\gamma) \leq \theta] \leq \gamma.$$

From

$$P[\Lambda(\beta) \geq \lambda] = P[c_1\Lambda(\beta) \geq c_1\lambda] \geq P[c_1\Lambda(\beta) \geq \mu]$$

and

$$P[\Theta(\gamma) \leq \theta] = P[c_2\Theta(\gamma) \leq c_2\theta] \geq P[c_2\Theta(\gamma) \leq \mu],$$

one clearly has that $(c_1\Lambda(\beta),c_2\Theta(\gamma))$ is a $1-\beta-\gamma$ confidence interval for μ.

Similarly, one can calculate an alternative lower confidence point for μ by using the inequality

$$\mu \geq c_1 \frac{E(\phi_i\psi_i V_i)}{E(\phi_i W_i)}$$

and the pairs $(\phi_i\psi_i V_i,\phi_i V_i)$ as well as another upper confidence point for μ by using the inequality

$$\mu \leq c_2 \frac{E(\phi_i\psi_i W_i)}{E(\phi_i W_i)}$$

and the pairs $(\phi_i\psi_i W_i,\phi_i W_i)$.

5. Examples

We consider two models, a network with random arc lengths and a stationary (s,S) inventory system with periodic review. Figure 1 depicts a network with 5 arcs, source node $v_s = 1$, and terminal node $v_t = 4$. The arcs have independent, discrete random lengths with distributions listed in Table 1.

Insert Figure 1 and Table 1 here

Let B_i denote the length of arc i and let $L_P(\mathbf{B})$ denote the length of a path P from 1 to 4, where $\mathbf{B} = (B_1,\ldots,B_5)$. The length of a shortest path is then given by $L(\mathbf{B}) = \min_P L_P(\mathbf{B})$. We consider the estimation of the probability that the path $P_0 = \{1,4\}$ is shortest when $L(\mathbf{B}) \leq 5$ given by $\mu = P(L_{P_0}(\mathbf{B}) = L(\mathbf{B}) \,|\, L(\mathbf{B}) \leq 5) = 0.1818$. This probability is estimated by means of a crude Monte Carlo experiment with n (independent) trials. In trial i, we generate a sample $\mathbf{b}^{(i)} = (b_1^{(i)},\ldots,b_5^{(i)})$ and set $\phi_i = 1(L(\mathbf{b}^{(i)}) \leq 5)$, $\psi_i = 1(L_{P_0}(\mathbf{b}^{(i)}) = L(\mathbf{b}^{(i)}))$, and $W_i = 1$.

For each combination of error α, sample size n, and upper bound m, the coverage for each class of confidence intervals was estimated by the proportion of confidence intervals resulting from 200 independent replications that contained the true parameter μ. The new confidence intervals were computed with the heuristic assignment $\delta = \bar{Y}\alpha/(\bar{\mu}+\bar{Y})$, $\beta = \bar{\mu}(\alpha-\delta)$ and $\gamma = \alpha-\beta-\delta$ and were roughly 4 percent narrower than those computed with $\delta = \alpha/2$ and $\beta = \gamma = \alpha/4$.

Insert Table 2 here

The binomial confidence intervals were computed with $\beta = \gamma = \alpha/2$ and the routine in Press *et al.* (1989, pp. 166-168) was used for evaluating the inverse Beta distribution function. These intervals are conservative with their average widths being 1.23 times larger (on the average) than those of the classical intervals but 1.47 times smaller than those of the proposed intervals. As expected, for fixed α and for $n \geq 50$ the average width decreases roughly with $\sqrt{n}$. It should be mentioned that for $n \geq 20$ numerous underflows occurred during the computation of the inverse Beta distribution.

The remaining results agree with our observations in Sections 2 and 3. Indeed, note

that for $n \leq 25$ the proportion of jackknife confidence intervals that contain μ is considerably smaller than the desired level 1-α. For instance, when $\alpha = 0.05$ and $n = 25$ the jackknife interval has estimated coverage 0.830 which is considerably below the intended 0.95. Also note that the new intervals are conservative since they have coverage close to one.

Now observe that for fixed α and m and for $n \geq 50$, the width of the proposed intervals decreases roughly with $1/\sqrt{n}$. In fact, the average width of the proposed interval decreases at a rate that is comparable with that for the jackknife interval (consider for example the cases $\alpha = 0.05$, $n = 50$ and $\alpha = 0.05$, $n = 100$).

We now discuss the growth of the width of the proposed interval with the upper bound m. Observe that for fixed α and $n \geq 50$ this width grows proportionally to $\sqrt{m}$. For instance, when $\alpha = 0.10$ and $n = 100$ increasing m from its exact value $m = 1$ to 2 and 4 results in an increase of the average width by factors of $0.3514/0.2486 = 1.41 \simeq \sqrt{2}$ and $0.4930/0.2486 = 1.98$. This increase is in line with the claim made at the beginning of Section 3.

The second example considers the probability that the path P_0 is shortest given that $L(\mathbf{B}) \leq 5$ when $P(B_1 = 2) = 0.6$ and $P(B_1 = 3) = 0.4$ while the distributions of the remaining arc lengths are as in Table 1.

Insert Table 3 here

The results in Table 3 were obtained by using the samples from the first experiment and the factors $W_i = (0.6/0.5)1(b_1^{(i)} = 2) + (0.4/0.5)1(b_1^{(i)} = 3) \leq 1.2$. These results are in agreement with those in Table 2 as for sample sizes $n \leq 50$ the classical and jackknife intervals exhibit low coverage while for larger sample sizes the latter intervals have roughly equal average widths and coverages.

The second model is an inventory system with periodic review operating under a

stationary (s,S) policy with $s = 3$ and $S = 6$ (see Heyman and Sobel 1982, example 7-4). Let D_k denote the demand during period k. We assume that D_k, $k = 1,2,...$ are i.i.d. random variables with distribution $P(D_k = 1) = 0.10$, $P(D_k = 2) = 0.90$. Let Z_k be the inventory on hand plus on order at the beginning of period k before demand occurs. Then $\{Z_k: k \geq 1\}$ is a discrete time Markov process with four states and its limiting distribution can be easily computed (Heyman and Sobel 1982, example 7-11). We want to estimate $\lim_{k\to\infty} P(Z_k \geq 4) = 0.9174$. Since $\{Z_k: k \geq 1\}$ is also a regenerative process, the latter probability can also be estimated by simulating the system over n cycles and collecting X_i as the time during the ith cycle when $Z_k \geq 4$ and Y_i as the ith cycle time. We used S as the regenerative state and 4 as the upper bound on both X_i and Y_i.

Insert Table 4 here

Table 4 contains the experimental results. The confidence interval from Corollary 1 in column 4 is roughly five times wider than the classical or the jackknife confidence intervals but has superior coverage for sample sizes up to 100. The classical and jackknife interval estimates are narrow because the $\text{var}(X_i - \mu Y_i)$ is small. On the other hand, the bounds in Theorem 1 (ii)-(iii) depend only on the means μ_x and μ_y. Note that the average widths in column 4 are comparable in size with the widths in column 4 of Tables 2 and 3. As in the previous two experiments, for $n \geq 50$ the average width of the new interval decreases with $1/\sqrt{n}$ for fixed α. Finally, when the right endpoint 4 of the c.d.f. of Y_i was replaced by 8, the average width of the proposed interval for $n = 100$ increased approximately by a factor of 2.

6. Conclusions and Recommendations

We have proposed confidence intervals for the ratio $\mu = \mu_x/\mu_y$ of two means estimated by using i.i.d. random pairs (X_i, Y_i) with $0 \leq X_i \leq Y_i \leq 1$. These intervals were obtained by

using distribution-free bounds on the probability $P(\overline{X}-\mu\overline{Y} \geq \epsilon)$. The bounds resulted from an inequality due to S.N. Bernstein and are tight. The interval estimators have coverage close to unity regardless of the sample size and the nominal coverage; and these interval estimators can be used in place of confidence intervals based on the central limit theorem, especially for small sample sizes (≤ 50).

We considered two cases. In the basic case, the random variables X_i and Y_i can be expressed as $X_i = \phi_i \psi_i W_i$ and $X_i = \phi_i W_i$, where ϕ_i and ψ_i are Bernoulli r.v.'s and W_i is also a r.v. Random variables of this type appear frequently in Monte Carlo simulations of stochastic networks. In the general case, the variables X_i and Y_i follow arbitrary distributions and the proposed confidence interval can be used for estimating limiting probabilities of regenerative processes with bounded cycle lengths. If the upper bound on Y_i is $m \neq 1$, the the confidence intervals can be computed by using the r.v.'s $X_i^* = X_i/m$ and $Y_i^* = Y_i/m$. This transformation does not affect the resulting intervals as long as m is the right endpoint of the c.d.f. of Y_i. If m is an arbitrary bound, then the confidence interval in Theorem 3 for the basic case appears to widen proportionally to $\sqrt{m}$ while the confidence interval in Corollary 1 appears to widen proportionally to m.

Overall, the confidence interval for the basic case outperformed the interval for the general case with regard to the ratio of its width over the width of the jackknife (or classical) confidence interval. This performance is due to the small $\text{var}(X_i - \mu Y_i)$ in the latter case which is not accounted for by the bounds in Theorem 1 (ii)-(iii).

Narrower confidence intervals can be obtained by using tighter bounds for the probabilities $P[\overline{X}-\mu\overline{Y} \geq \epsilon]$ and $P[-\overline{X}+\mu\overline{Y} \geq \epsilon]$. One such bound is due to Hoeffding (1963, inequality (2.8))

$$P[\overline{X}-\mu\overline{Y} \geq \epsilon] \leq (1+b\epsilon/\sigma^2)^{-(b\epsilon+\sigma^2)/(b^2+\sigma^2)}(1-\epsilon/b)^{-(b^2-\epsilon)/(b^2+\sigma^2)},$$

where b is an upper bound on $X_i-\mu Y_i$ (if $0 \leq X_i \leq Y_i \leq 1$, b can be chosen as $1-\mu$) and $\sigma^2 = \text{var}(X_i-\mu Y_i)$. A similar bound can be obtained for the probability $P[-\overline{X}+\mu\overline{Y} \geq \epsilon]$. Unfortunately, these bounds contain the unknown variance σ^2. If X_i and Y_i are Bernoulli r.v.'s, $\sigma^2 = \mu_x(1-\mu)$ but in other cases σ^2 must be estimated. The use of inequalities containing higher moments of X_i and Y_i is a problem worth future consideration.[1]

[1] I would like to thank Professor George Fishman at the University of North Carolina for our discussions and Professor James Wilson at North Carolina State University for his careful reading of the manuscript and several valuable suggestions. Many thanks to an anonymous reviewer for suggesting several stylistic changes.

References

ABRAMOWITZ, M. AND I. STEGUN, *Handbook of Mathematical Functions* (Applied Mathematics Series 55), National Bureau of Standards, Washington, D.C., 1964.

ALEXOPOULOS, C. AND G. S. FISHMAN, "Characterizing Stochastic Flow Networks Using the Monte Carlo Method," *Networks*, 21 (1991), 775-978.

BAZARAA, M. S. AND C. M. SHETTY, *Nonlinear Programming: Theory and Algorithms*, John Wiley and Sons, 1979.

FIELLER, E. C., "Some Problems in Interval Estimation," *Journal of Royal, Statistical Society, Series B*, 2 (1954), 175-185.

FISHMAN, G. S., "Maximal Flow and Critical Cutset as Descriptors of Multi-state Systems with Randomly Capacitated Components," *Computers and Operations Research*, 14 (1987), 507-520.

———, "Monte Carlo, Control Variates and Stochastic Ordering," *SIAM Journal on Scientific and Statistical Computing*, 10 (1989), 187-204.

———, "Confidence Intervals for the Mean in the Bounded Case," *Statistics and Probability Letters*, 12 (1991), 223-227.

HEYMAN, D. P. AND M. J. SOBEL, *Stochastic Models in Operations Research: Volume I*, McGraw-Hill, 1982.

HOEFFDING, W., "Probability Inequalities for Sums of Bounded Random Variables," *American Statistical Association Journal*, 58 (1963), 13-29.

IGLEHART, D. L., "Simulating Stable Stochastic Systems, V: Comparison of Ratio Estimators," *Naval Research Logistics Quarterly*, 22 (1975), 553-565.

IMSL, Inc., *IMSL Library Reference Manual*, Ed. 8, Houston, Texas, 1982.

MARSDEN, J. E., *Elementary Classical Analysis*, W. H. Freeman and Company, 1974.

PRESS, W. H., B. P. FLANNERY, S. A. TEUKOLSKY, AND W. T. VETTERLING, *Numerical Recipes: The art of Scientific Computing (FORTRAN version)*, Cambridge University Press, 1989.

SHIRYAYEV, A. N., *Probability*, Springer-Verlag, 1984.

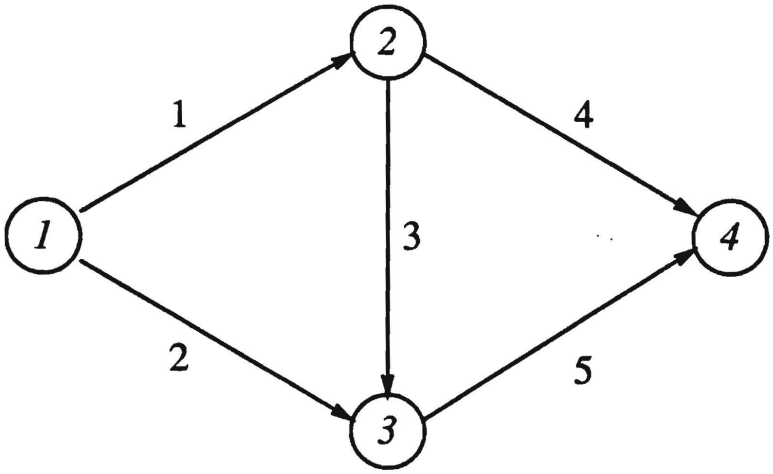

Figure 1

Table 1

Distributions of arc lengths in Figure 1

Arc	Lengths	Probabilities
(1,2)	2, 3	0.5, 0.5
(1,3)	1, 2	0.5, 0.5
(2,3)	1, 3	0.5, 0.5
(2,4)	3, 4	0.5, 0.5
(3,4)	2, 4	0.5, 0.5

Table 2

Confidence intervals for the conditional probability that the path {1,4} is shortest given that the shortest path length is ≤ 5

			Average width†			Estimated coverage†		
α	n	m	New	Binomial	Jackknife	New	Binomial	Jackknife
0.10	10	1	.6178	.4506	.2999	.955	.970	.655
	15	1	.5660	.3756	.2903	.985	.970	.820
	20	1	.5169	.3273	.2709	1.0	0.940	.870
	25	1	.4732	.2907	.2480	.995	.955	.805
	50	1	.3394	.1978	.1777	.995	.915	.865
		2	.4936					
		4	.6825					
	100	1	.2486	.1399	.1296	1.0	.930	.890
		2	.3514					
		4	.4930					
0.05	10	1	.6598	.5112	.3374	.955	.975	.670
	15	1	.6054	.4295	.3318	1.0	.985	.830
	20	1	.5532	.3758	.3113	1.0	.980	.915
	25	1	.5068	.3351	.2878	1.0	.985	.830
	50	1	.3640	.2303	.2106	.995	.960	.915
		2	.5375					
		4	.7365					
	100	1	.2664	.1641	.1545	1.0	.980	.940
		2	.3830					
		4	.5425					

† Estimated from 200 independent experiments.

Table 2 (continued)

Confidence intervals for the conditional probability that the path {1,4} is shortest given that the shortest path length is ≤ 5

			Average width†			Estimated coverage†		
α	n	m	New	Binomial	Jackknife	New	Binomial	Jackknife
0.01	10	1	.7336	.6209	.4016	.990	1.0	.685
	15	1	.6792	.5294	.3990	0.990	.990	.840
	20	1	.6234	.4669	.3764	1.0	.995	.930
	25	1	.5733	.4192	.3520	1.0	1.0	.965
	50	1	.4152	.2927	.2686	1.0	.995	.940
		2	.6226					
		4	.8243					
	100	1	.3036	.2108	.2028	1.0	.990	.985
		2	.4486					
		4	.6375					

† Estimated from 200 independent experiments.

Table 3

Arc 1 has lengths 2 or 3 with respective probabilities 0.6 or 0.4
Samples were drawn with probabilities in Table 1

			Average width†			Estimated coverage†		
α	n	m	New	Classical	Jackknife	New	Classical	Jackknife
0.10	10	1.2	.7219	.3571	.3334	1.0	.705	.710
	15	1.2	.6545	.3330	.3205	1.0	.865	.865
	20	1.2	.5939	.2990	.2929	1.0	0.815	.815
	25	1.2	.5427	.2696	.2672	1.0	.880	.885
	50	1.2 2	.3928 .5092	.1901	.1901	1.0	.850	.850
	100	1.2 2	.2852 .3659	.1372	.1375	1.0	.860	.870
0.05	10	1.2	.7653	.4255	.3777	1.0	.725	.740
	15	1.2	.6965	.3968	.3679	1.0	.885	.885
	20	1.2	.6346	.3563	.3383	1.0	.910	.910
	25	1.2	.5816	.3213	.3117	1.0	.895	.895
	50	1.2 2	.4230 .5527	.2265	.2266	1.0	.900	.900
	100	1 2	.3071 .3983	.1635	.1638	1.0	.940	.940
0.01	10	1.2	.8365	.5592	.4490	1.0	.740	.740
	15	1.2	.7706	.5215	.4438	1.0	.895	.900
	20	1.2	.7095	.4682	.4128	1.0	.960	.960
	25	1.2	.6556	.4222	.3840	1.0	.915	.915
	50	1.2 2	.4845 .6364	.2977	.2928	1.0	.950	.950
	100	1 2	.3527 .4646	.2148	.2153	1.0	.995	.995

† Estimated from 200 independent experiments.

Table 4

(s,S) inventory system with $s = 3$ and $S = 6$
Estimation of $\lim_{k\to\infty} P(Z_k \geq 4)$

			Average width†			Estimated coverage†		
α	n	m	New	Classical	Jackknife	New	Classical	Jackknife
0.10	10	4	.7495	.1419	.1386	1.0	.795	.795
	15	4	.6480	.1262	.1243	1.0	.865	.890
	20	4	.5743	.1131	.1126	1.0	.845	.845
	25	4	.5178	.1010	.1012	1.0	.775	.775
	50	4	.3665	.0733	.0738	1.0	.870	.850
	100	4	.2555	.0525	.0526	1.0	.855	.855
		8	.5243			1.0		
0.05	10	4	.7986	.1691	.1570	1.0	.795	.795
	15	4	.7014	.1504	.1438	1.0	.930	.930
	20	4	.6278	.1347	.1319	1.0	.860	.860
	25	4	.5700	.1203	.1193	1.0	.905	.905
	50	4	.4085	.0874	.0878	1.0	.925	.925
	100	4	.2855	.0626	.0626	1.0	.920	.920
		8	.5734			1.0		
0.01	10	4	.8227	.2222	.1898	1.0	.820	.820
	15	4	.7893	.1971	.1763	1.0	.940	.940
	20	4	.7204	.1771	.1632	1.0	.975	.980
	25	4	.6634	.1581	.1500	1.0	.935	.935
	50	4	.4897	.1148	.1148	1.0	.970	.970
	100	4	.3458	.0822	.0825	1.0	.990	.990
		8	.6593			1.0		

† Estimated from 200 independent experiments.

Appendix

Proof of Theorem 1. (*i*) Fix $(t,r,\epsilon) \in S$ and note that

$$\frac{\partial F}{\partial h} = -\epsilon(1-t)e^{-\epsilon h} - t(1-r)(r+\epsilon)e^{-(r+\epsilon)h} + tr(1-r-\epsilon)e^{(1-r-\epsilon)h}$$

and

$$\frac{\partial^2 F}{\partial h^2} = \epsilon^2(1-t)e^{-\epsilon h} + t(1-r)(r+\epsilon)^2e^{-(r+\epsilon)h} + tr(1-r-\epsilon)^2e^{(1-r-\epsilon)h} > 0.$$

The strict convexity of F with respect to h along with the properties $\left.\frac{\partial F}{\partial h}\right|_{h=0} = -\epsilon < 0$ and $\lim_{h\to+\infty} \frac{\partial F}{\partial h} = +\infty$ assert the existence of a unique minimum $h^0 \in (0,\infty)$.

Since F is smooth and for each $(t,r,\epsilon) \in \text{int}(S)$ the system

$$\frac{\partial F}{\partial h}(t,h,r,\epsilon) = 0$$
$$(t,r,\epsilon) \in \text{int}(S),\ h \in (0,\infty)$$

has a solution, namely (t,h^0,r,ϵ), the Implicit Function Theorem (Marsden 1974, p. 210-211) implies the existence of a unique smooth function

$$h^*(t,r,\epsilon)\colon \text{int}(S) \longrightarrow (0,\infty)$$

that satisfies

$$\frac{\partial F}{\partial h}(t,h^*(t,r,\epsilon),r,\epsilon) = 0 \quad \text{for all } (t,r,\epsilon).$$

In addition, $F(t,h^*(t,r,\epsilon),r,\epsilon)$ is smooth in $\text{int}(S)$.

If $t = 1$, then $F(t,h,r,\epsilon)$ is minimized by

$$h^0 = \log\left[\frac{(1-r)(r+\epsilon)}{r(1-r-\epsilon)}\right]$$

and

$$F(1,h^0,r,\epsilon) = \exp\{G(r,\epsilon)\}.$$

(*iii*) We can easily verify that e^{hX_1} can be written as

$$e^{hX_1} = 1-\phi_1\psi_1+\phi_1\psi_1 e^{h\phi_1 W_1}$$

and then

$$e^{h(X_1-\mu Y_1)} = (1-\phi_1\psi_1)e^{-h\mu Y_1}+\phi_1\psi_1 e^{h(1-\mu)Y_1}.$$

For each realization $y \in [0,1]$ of Y_1, Jensen's inequality implies $e^{-h\mu y} \le 1-y+ye^{-h\mu}$ and $e^{h(1-\mu)y} \le 1-y+ye^{h(1-\mu)}$. Hence,

$$\begin{aligned} E\,e^{h(X_1-\mu Y_1)} &\le E\left\{(1-\phi_1\psi_1)\left[1-Y_1+Y_1e^{-h\mu}\right] + \phi_1\psi_1\left[1-Y_1+Y_1e^{h(1-\mu)}\right]\right\} \\ &= E\left\{1-Y_1+[Y_1-\phi_1\psi_1 W_1]e^{-h\mu} + \phi_1\psi_1 W_1 e^{h(1-\mu)}\right\} \\ &= 1-\mu_y+\mu_y\left[(1-\mu)\,e^{-h\mu} + \mu\,e^{h(1-\mu)}\right] \end{aligned}$$

and

$$P\,[\overline{X}-\mu\overline{Y} \ge \epsilon] \le \left[\min_{h\ge 0} F(\mu_y,h,\mu,\epsilon)\right]^n = [F(\mu_y,h^*(\mu_y,\mu,\epsilon),\mu,\epsilon)]^n.$$

The proof of (10) follows similarly if we rewrite $-\overline{X}+\mu\overline{Y}$ as $(\overline{Y}-\overline{X})-(1-\mu)\overline{Y}$ and note that $Y_i-X_i = \phi_i(1-\psi_i)W_i$. □

Proof of Theorem 2. (*i*) We first fix β and show that for each $0 < t < 1$ and each $0 < u \le v \le 1$ system (11) has a unique solution. Note that at $r = 0$ ($\epsilon = u > 0$) $F(t,h,0,u)$ is strictly decreasing in h and

$$\lim_{h\to+\infty} F(t,h,0,u) = 0 < \beta^{1/n} \tag{A.1}$$

while at $r = u/v$ ($\epsilon = 0$)

$$\min_{h\ge 0} F(t,h,u/v,0) = F(t,0,u/v,0) = 1 > \beta^{1/n}. \tag{A.2}$$

Let

$$h^* = h^*(t,r,\epsilon),\ \epsilon^* = \epsilon^*(r) = u\text{-}rv,\ F(h^*,\epsilon^*) = F(t,h^*(t,r,\epsilon^*(r)),r,\epsilon^*(r)),$$

and

$$\frac{\partial F}{\partial z}(h^*,\epsilon^*) = \frac{\partial F}{\partial z}\Big|_{h=h^*,\epsilon=\epsilon^*} \quad \text{for } z = h,\ r,\ \epsilon.$$

Any solution to (11) satisfies

$$\frac{dF}{dr}(h^*,\epsilon^*) = \frac{\partial F}{\partial r}(h^*,\epsilon^*) + \frac{d\epsilon^*}{dr}\frac{\partial F}{\partial \epsilon}\Big|_{h=h^*,\epsilon=\epsilon^*} =$$

$$\frac{\partial F}{\partial r}(h^*,\epsilon^*) - v\frac{\partial F}{\partial \epsilon}(h^*,\epsilon^*) = \frac{\partial F}{\partial r}(h^*,\epsilon^*) + vh^*F(h^*,\epsilon^*)$$

because $\frac{d\epsilon^*}{dr} = -v$ and

$$\frac{\partial F}{\partial \epsilon}(t,h,r,\epsilon) = -hF(t,h,r,\epsilon) < 0 \quad \text{for all } (t,h,r,\epsilon).$$

Also, $\frac{\partial F}{\partial h}(h^*,\epsilon^*) = 0$ implies

$$F(h^*,\epsilon^*) = tr(1\text{-}r)e^{-(r+\epsilon^*)h^*}(e^{h^*}\text{-}1)/\epsilon^*$$

and then

$$\frac{dF}{dr}(h^*,\epsilon^*) = te^{-(r+\epsilon^*)h^*}\{e^{h^*}\text{-}1\text{-}h^*+h^*r(e^{h^*}\text{-}1)(v\frac{1\text{-}r}{\epsilon^*}\text{-}1)\} > 0$$

because $e^x\text{-}1\text{-}x > 0$, $x > 0$ and $\epsilon^* = u\text{-}vr < v(1\text{-}r)$.

The fact that $F(t,h^*(t,r,u\text{-}vr),r,u\text{-}vr)$ is strictly increasing in $r \in (0,u/v)$ along with (A.1) and (A.2) imply that the system (11) has a unique solution $r'_\beta(t,u,v)$. Let $r' = r'_\beta(t,u,v)$ and $\epsilon' = 1\text{-}r'$. Note that $F(t,h,r',1\text{-}r') = (1\text{-}t)e^{-(1-r')h}+tr'+t(1\text{-}r')e^{-h}$ is strictly decreasing in h with $F(t,0,r',1\text{-}r') = 1 > \beta^{1/n}$ and $\lim_{h\to\infty} F(t,h,r'1\text{-}r') = tr'$. Therefore the solution to system (11) exists only if $tr' < \beta^{1/n}$. This and the inequality $r' < u/v$ yield $r' < \min\{u/v,\beta^{1/n}/t\}$.

Now choose a pair $(t_0,r_0) \in R$. The fact that for $\epsilon = 1\text{-}r_0$, $F(t_0,h,r_0,\epsilon) =$

$(1-t_0)e^{-(1-r_0)h}+t_0r_0+t_0(1-r_0)e^{-h}$ is strictly decreasing in h and approaches t_0r_0 as $h \to +\infty$ along with $\min_{h\geq 0} F(t_0,h,r_0,0) = 1 > \beta^{1/n}$ and the monotonicity of the smooth function $F(t,h^*(t,r,\epsilon),r,\epsilon)$ on the set $\{(t_0,r_0,\epsilon): 0 < \epsilon < 1-r_0\}$ imply the existence of an $\epsilon_0 \in (0,1-r_0)$ such that $(t_0,h^*(t_0,r_0,\epsilon_0),r_0,\epsilon_0)$ solves

$$\begin{aligned} F(t,h,r,\epsilon) &= \beta^{1/n} \\ \frac{\partial F}{\partial h}(t,h,r,\epsilon) &= 0 \\ (t,r,\epsilon) \in \text{int}(S),\ tr < \beta^{1/n},\ h &\in (0,\infty). \end{aligned} \tag{A.3}$$

Since any solution to (A.3) satisfies

$$\begin{vmatrix} \frac{\partial F}{\partial h} & \frac{\partial F}{\partial \epsilon} \\ \frac{\partial^2 F}{\partial h^2} & \frac{\partial^2 F}{\partial h \partial \epsilon} \end{vmatrix} = hF(t,h,r,\epsilon)\frac{\partial^2 F}{\partial h^2}(t,h,r,\epsilon) \neq 0,$$

the Implicit Function Theorem asserts the existence of an open neighborhood A of (t_0,r_0) and unique smooth functions $\bar{h}(t,r)$: $A \longrightarrow (0,\infty)$, $\bar{\epsilon}(t,r)$: $A \longrightarrow (0,1)$ such that $\bar{h}(t_0,r_0) = h^*(t_0,r_0,\epsilon_0)$, $\bar{\epsilon}(t_0,r_0) = \epsilon_0$ and

$$\begin{aligned} F(t,\bar{h}(t,r),r,\bar{\epsilon}(t,r)) &= \beta^{1/n} \\ \frac{\partial F}{\partial h}(t,\bar{h}(t,r),r,\bar{\epsilon}(t,r)) &= 0 \quad \text{for all } (t,r) \in A \\ \bar{\epsilon}(t,r) &< 1-r\,. \end{aligned} \tag{A.4}$$

Now Theorem 1 and (A.4) imply $h^*(t,r,\bar{\epsilon}(t,r)) = \bar{h}(t,r)$ for all $(t,r) \in A$.

The uniqueness of the functions $\bar{h}(t,r)$ and $\bar{\epsilon}(t,r)$ in a neighborhood of each $(t_0,r_0) \in R$ implies the existence and uniqueness of smooth functions

$$\bar{h}(t,r)\text{: } R \longrightarrow (0,\infty)$$

$$\bar{\epsilon}(t,r)\colon R \longrightarrow (0,1)$$

which satisfy (A.4) and $h^*(t,r,\bar{\epsilon}(t,r)) = \bar{h}(t,r)$ for all (t,r).

We now show that for fixed $0 < t < 1$ and $0 < u \leq v \leq 1$, the curve $\bar{\epsilon}(t,r)$ and the line $\epsilon^*(r) = u - rv$ have a unique intersection, namely at $r = r'_\beta(t,u,v)$. Indeed, the point $(t,r'_\beta(t,u,v),u-r'_\beta(t,u,v)\cdot v)$ solves systems (11) and (A.3). Then, there exist unique $\bar{h}(t,r'_\beta(t,u,v))$ and $\bar{\epsilon}(t,r'_\beta(t,u,v))$ such that $\bar{\epsilon}(t,r'_\beta(t,u,v)) = u - r'_\beta(t,u,v)\cdot v$, $\bar{h}(t,r'_\beta(t,u,v)) = h^*(t,r'_\beta(t,u,v),\bar{\epsilon}(t,r'_\beta(t,u,v)))$, and $(t,r'_\beta(t,u,v),\bar{\epsilon}(t,r'_\beta(t,u,v)))$ satisfies (A.4).

Showing that $\{\bar{\epsilon}(t,r)\}$ is a family of strongly quasiconcave curves indexed by $t \in (0,1)$

We now fix t, let $b_t = \min\{1,\beta^{1/n}/t\}$, and suppress the functional dependence of $\bar{h}(t,r)$ and $\bar{\epsilon}(t,r)$ on t. We also use $\partial^k F(\bar{h})/\partial r^i \partial \epsilon^j$ to denote the partial derivative $\partial^k F/\partial r^i \partial \epsilon^j$ evaluated at $(t,\bar{h}(r),r,\bar{\epsilon}(r))$.

We have

$$0 \equiv \frac{dF}{dr}(\bar{h}) = \frac{\partial F}{\partial r}(\bar{h}) + \frac{\partial F}{\partial h}(\bar{h})\frac{d\bar{h}}{dr} + \frac{\partial F}{\partial \epsilon}(\bar{h})\frac{d\bar{\epsilon}}{dr} = \frac{\partial F}{\partial r}(\bar{h}) + \frac{\partial F}{\partial \epsilon}(\bar{h})\frac{d\bar{\epsilon}}{dr}$$

implying

$$\frac{d\bar{\epsilon}}{dr} = -\frac{\partial F}{\partial r}(\bar{h}) \,/\, \frac{\partial F}{\partial \epsilon}(\bar{h}). \tag{A.5}$$

Differentiation of (A.5) that uses the identities $dF(\bar{h})/dr \equiv 0$ and $\partial F(\bar{h})/\partial \epsilon = -\bar{h}(r)F(\bar{h})$ yields

$$\frac{d^2\bar{\epsilon}}{dr^2} = \left\{\frac{\partial^2 F}{\partial r^2}(\bar{h}) + \frac{\partial^2 F}{\partial h \partial r}(\bar{h})\frac{d\bar{h}}{dr} + \frac{\partial^2 F}{\partial \epsilon \partial r}(\bar{h})\frac{d\bar{\epsilon}}{dr} - F(\bar{h})\frac{d\bar{\epsilon}}{dr}\frac{d\bar{h}}{dr}\right\}/[\bar{h}(r)F(\bar{h})]. \tag{A.6}$$

Since $(t,\bar{h}(r),r,\bar{\epsilon}(r))$ satisfy $\frac{\partial F}{\partial h}(t,\bar{h}(t,r),r,\bar{\epsilon}(t,r)) = 0$ for all $r \in (0,b_t)$, we have

$$0 = \frac{d}{dr}\left[\frac{\partial F}{\partial h}(\bar{h})\right] = \frac{\partial^2 F}{\partial h^2}(\bar{h})\frac{d\bar{h}}{dr} + \frac{\partial^2 F}{\partial r \partial h}(\bar{h}) + \frac{\partial^2 F}{\partial \epsilon \partial h}(\bar{h})\frac{d\bar{\epsilon}}{dr}$$

and then

$$\frac{d\bar{h}}{dr} = -\left\{\frac{\partial^2 F}{\partial r \partial h}(\bar{h}) + \frac{\partial^2 F}{\partial \epsilon \partial h}(\bar{h})\,\frac{d\bar{\epsilon}}{dr}\right\}\Big/\frac{\partial^2 F}{\partial h^2}(\bar{h}). \tag{A.7}$$

Substitution of (A.7) into (A.6) gives

$$\begin{aligned}\frac{d^2\bar{\epsilon}}{dr^2} = \Bigg\{&\frac{\partial^2 F}{\partial r^2}(\bar{h}) - \frac{\partial^2 F}{\partial h \partial r}(\bar{h})\left[\frac{\partial^2 F}{\partial h \partial r}(\bar{h}) + \frac{\partial^2 F}{\partial \epsilon \partial h}(\bar{h})\,\frac{d\bar{\epsilon}}{dr}\right]\Big/\frac{\partial^2 F}{\partial h^2}(\bar{h}) + \frac{\partial^2 F}{\partial \epsilon \partial r}(\bar{h})\,\frac{d\bar{\epsilon}}{dr} \\ &+ F(\bar{h})\,\frac{d\bar{\epsilon}}{dr}\left[\frac{\partial^2 F}{\partial h \partial r}(\bar{h}) + \frac{\partial^2 F}{\partial \epsilon \partial h}(\bar{h})\,\frac{d\bar{\epsilon}}{dr}\right]\Big/\frac{\partial^2 F}{\partial h^2}(\bar{h})\Bigg\}\Big/[\bar{h}(r)F(\bar{h})].\end{aligned} \tag{A.8}$$

Suppose now that there is an $r_0 \in (0,b_t)$ for which $d\bar{\epsilon}(r_0)/dr = 0$. At $r = r_0$, (A.8) gives

$$\frac{d^2\bar{\epsilon}}{dr^2} = \left\{\frac{\partial^2 F}{\partial r^2}(\bar{h})\,\frac{\partial^2 F}{\partial h^2}(\bar{h}) - \left[\frac{\partial^2 F}{\partial h \partial r}(\bar{h})\right]^2\right\}\Big/\left[\bar{h}(r)F(\bar{h})\,\frac{\partial^2 F}{\partial h^2}(\bar{h})\right] \tag{A.9}$$

and (A.5) implies

$$\frac{\partial F}{\partial r}(\bar{h}) = -\frac{\partial F}{\partial \epsilon}(\bar{h})\,\frac{d\bar{\epsilon}}{dr} = 0. \tag{A.10}$$

Since

$$\frac{\partial F}{\partial r}(\bar{h}) = -\bar{h}(r)F(\bar{h}) + e^{-\bar{\epsilon}(r)\bar{h}(r)}\{(1-t)\bar{h}(r) + t\,e^{-r\bar{h}(r)}(e^{\bar{h}(r)}-1)\},$$

(A.10) implies that at $r = r_0$

$$\frac{\partial^2 F}{\partial r^2}(\bar{h}) = -t\bar{h}(r)\,e^{-(r+\bar{\epsilon}(r))\bar{h}(r)}(e^{\bar{h}(r)}-1) < 0.$$

Finally, the inequality $\partial^2 F(\bar{h})/\partial h^2 > 0$ and (A.9) imply

$$d\bar{\epsilon}^2(r)/dr^2 < 0 \text{ for all } r \in (0,b_t) \text{ such that } d\bar{\epsilon}(r)/dr = 0 \tag{A.11}$$

and then any stationary point of $\bar{\epsilon}(r)$ is a local maximum.

We now prove that $\bar{\epsilon}(r)$, $r \in (0,b_t)$ is strongly quasiconcave, that is, for all $0 < r_1 < r_2 < b_t$, $\bar{\epsilon}(r) > \min\{\bar{\epsilon}(r_1),\bar{\epsilon}(r_2)\}$ for all $r \in (r_1,r_2)$ (see Bazaraa and Shetty 1979, pp. 99-111). If $\bar{\epsilon}(\cdot)$ is not strongly quasiconcave, then there is an $\hat{r} \in (r_1,r_2)$ such that

$$\bar{\epsilon}(\hat{r}) \leq \min\{\bar{\epsilon}(r_1),\bar{\epsilon}(r_2)\}. \tag{A.12}$$

Now we seek an $\tilde{r} \in (r_1,r_2)$ such that

$$\bar{\epsilon}(\tilde{r}) < \min\{\bar{\epsilon}(r_1),\bar{\epsilon}(r_2)\}. \tag{A.13}$$

If (A.12) is a strict inequality, then we take $\tilde{r} = \hat{r}$; otherwise we proceed as follows to find $\tilde{r}$ in a sufficiently small neighborhood of $\hat{r}$. We have the following two cases:

Case 1: $d\bar{\epsilon}(\hat{r})/dr = 0$. In this case, we apply Taylor's formula with second-degree remainder and (A.11) to see that for every $\tilde{r} \neq \hat{r}$ in a sufficiently small neighborhood of $\hat{r}$,

$$r_1 < \tilde{r} < r_2 \quad \text{and} \quad \bar{\epsilon}(\tilde{r}) < \bar{\epsilon}(\hat{r}) \tag{A.14}$$

so that (A.13) holds.

Case 2: $d\bar{\epsilon}(\hat{r})/dr \neq 0$. In this case, we apply Taylor's formula with first-degree remainder to see that if $d\bar{\epsilon}(\hat{r})/dr > 0$ (respectively, $d\bar{\epsilon}(\hat{r})/dr < 0$), then for every $\tilde{r}$ in a sufficiently small open interval with $\hat{r}$ as the upper (respectively, lower) endpoint, (A.14) holds.

Since the function $\bar{\epsilon}(\cdot)$ is continuous on the compact interval $[r_1,r_2]$, there is a point $r^* \in [r_1,r_2]$ at which $\bar{\epsilon}(\cdot)$ attains its global minimum on $[r_1,r_2]$:

$$\bar{\epsilon}(r^*) = \inf\{\bar{\epsilon}(r): r \in [r_1,r_2]\}. \tag{A.15}$$

Now (A.13), (A.15), and the continuous differentiability of $\bar{\epsilon}(\cdot)$ imply

$$r^* \in (r_1, r_2) \quad \text{and} \quad d\bar{\epsilon}(r^*)/dr = 0. \tag{A.16}$$

However, (A.16), (A.11), and Taylor's formula with second-degree remainder imply that r^* must be a local maximum of $\bar{\epsilon}(\cdot)$; a contradiction to (A.15). It follows that $\bar{\epsilon}(\cdot)$ is strongly quasiconcave. We plotted $\bar{\epsilon}(\cdot)$ for several values of t, β and n. Although all the plots showed evidence that $\bar{\epsilon}(\cdot)$ is concave, we could not prove it.

Now consider a point $(r_0, \bar{\epsilon}(r_0))$, $r_0 \in (0, b_t)$ and let $0 < u_1 < \min\{r_0, \bar{\epsilon}(r_0)\}$. The intersection between the curve $\bar{\epsilon}(r)$ and the line $\epsilon = u_1 - r$ is a point $(r_1, \bar{\epsilon}(r_1))$ such that $r_1 < r_0$ and $\bar{\epsilon}(r_1) < \bar{\epsilon}(r_0)$. This argument implies that the origin (0,0) is an accumulation point of the set $\{(r, \bar{\epsilon}(r)),\ 0 < r < b_t\}$. If there are $0 < r_1 < r_2$ such that $\bar{\epsilon}(r_1) \geq \bar{\epsilon}(r_2)$, then $\bar{\epsilon}(r)$ has a local maximum and therefore a unique global maximum because $\bar{\epsilon}(\cdot)$ is strongly quasiconcave. Otherwise, $\bar{\epsilon}(\cdot)$ is strictly increasing. In both cases $\lim_{r \to 0^+} \bar{\epsilon}(r) = 0$.

(*ii*) The case $t = 1$ deserves special attention because of its simplicity. Since

$$f(r, \epsilon) = \min_{h \geq 0} F(1, h, r, \epsilon) = \exp\{G(r, \epsilon)\}$$

is smooth on $S_1 = \{(r, \epsilon):\ 0 < r < \beta^{1/n},\ 0 < \epsilon < 1 - r\}$ and

$$\frac{\partial f}{\partial \epsilon} = -\log\left[\frac{(1-r)(r+\epsilon)}{r(1-r-\epsilon)}\right] f(r, \epsilon) < 0 \quad \text{for all } (r, \epsilon) \in S_1,$$

an argument similar to that in part (i) asserts the existence and uniqueness of a smooth function $\hat{\epsilon}(r)$: $(0, \beta^{1/n}) \longrightarrow (0,1)$ which satisfies (13). In addition, $\hat{\epsilon}(r)$ can be easily shown

to be strictly concave with $\lim_{r\to 0^+} \hat{\epsilon}(r) = 0$. The proof of the existence of a solution to system (12) is a special case of that for system (11).

(*iii*) We first prove inequality (16) for $t = \mu_y$. Assume $\overline{X} > 0$ and note that $0 < L(\mu_y,\beta) < \min\{1,\beta^{1/n}/\mu_y\}$. Since for fixed $0 < u \le v \le 1$ the line $\epsilon^*(r) = u - rv$ has a unique intersection with the curve $\bar{\epsilon}(\mu_y,r)$, namely $r'_\beta(\mu_y,u,v)$, and $\lim_{r\to 0^+} \bar{\epsilon}(\mu_y,r) = 0 < u = \epsilon^*(0)$, it follows that $u - rv \ge \bar{\epsilon}(\mu_y,r)$ for all $r \le r'_\beta(\mu_y,u,v)$. Using $u = \overline{X}$, $v = \overline{Y}$ and $r'_\beta(\mu_y,u,v) = L(\mu_y,\beta)$, one has $\overline{X} - \mu\overline{Y} \ge \bar{\epsilon}(\mu_y,\mu)$ for $\mu \le L(\mu_y,\beta)$. Therefore

$$\begin{aligned} P\,[L(\mu_y,\beta) \ge \mu] &= P\,[L(\mu_y,\beta) \ge \mu, \overline{X} > 0] + P\,[L(\mu_y,\beta) \ge \mu, \overline{X} = 0] \\ &= P\,[L(\mu_y,\beta) \ge \mu, \overline{X} > 0] \le P\,[\overline{X} - \mu\overline{Y} \ge \bar{\epsilon}(\mu_y,\mu)] \\ &\le [F(\mu_y, h^*(\mu_y,\mu,\bar{\epsilon}(\mu_y,\mu)),\mu,\bar{\epsilon}(\mu_y,\mu))]^n \\ &= [F(\mu_y,\bar{h}(\mu_y,\mu),\mu,\bar{\epsilon}(\mu_y,\mu))]^n = \beta. \end{aligned}$$

An analogous argument shows that system (11) has a unique solution for $u = \overline{Y} - \overline{X}$ and $v = \overline{Y}$ and defines a lower confidence point for $1-\mu$ when $\overline{X} < \overline{Y}$. One then has $P\,[1 - U(\mu_y,\gamma) \ge 1-\mu] \le \gamma$ and hence $P\,[L(\mu_y,\beta) < \mu < U(\mu_y,\gamma)] \ge 1-\beta-\gamma = 1-\alpha$.

To show that the random variables $L(t,\beta)$ and $U(t,\gamma)$ are monotone in t w.p. 1, and hence complete the proof, it suffices to show that $\bar{\epsilon}(t_1,r) \le \bar{\epsilon}(t_2,r)$ for all $0 < t_1 < t_2 \le 1$ and $0 < t_1 r < t_2 r < \beta^{1/n}$. To show this we prove that for fixed (t,r), $\partial\bar{\epsilon}/\partial t \ge 0$. Evaluating derivatives at $(t,\bar{h}(t,r),r,\bar{\epsilon}(t,r))$ one has

$$0 \equiv \frac{dF}{dt} = \frac{\partial F}{\partial t} + \frac{\partial F}{\partial h}\frac{d\bar{h}}{dt} + \frac{\partial F}{\partial \epsilon}\frac{\partial \bar{\epsilon}}{\partial r} = \frac{\partial F}{\partial t} + \frac{\partial F}{\partial \epsilon}\frac{\partial \bar{\epsilon}}{\partial t}$$

implying that $\partial\bar{\epsilon}/\partial t = -\dfrac{\partial F/\partial t}{\partial F/\partial \epsilon}$. Since $\dfrac{\partial F}{\partial \epsilon}(t,\bar{h}(t,r),r,\bar{\epsilon}(t,r)) = -\bar{h}(t,r)F(t,\bar{h}(t,r),r,\bar{\epsilon}(t,r)) < 0$, it suffices to show that $\dfrac{\partial F}{\partial t}(t,h,r,\epsilon) \ge 0$ for all $(t,r,\epsilon) \in S$ and $h \in (0,\infty)$. Since $F(t,h,r,\epsilon)$ is

linear in t, we need to show that

$$-1+(1-r)e^{-rh}+r\,e^{(1-r)h} \geq 0 \Leftrightarrow e^{-rh}[1-r+r\,e^{h}-e^{rh}] \geq 0 \Leftrightarrow$$

$$1-r+r\,e^{h}-e^{rh} \geq 0 \Leftrightarrow 1-r+r\sum_{i=0}^{\infty} h^i/i! - \sum_{i=0}^{\infty} (rh)^i/i! \geq 0 \Leftrightarrow$$

$$r\sum_{i=1}^{\infty} h^i/i! - \sum_{i=1}^{\infty} (rh)^i/i! \geq 0 \quad \text{since } 0 < r < 1. \ \square$$

Proof of Theorem 4. Since $G(r,\epsilon) \leq -2\epsilon^2$ (see Hoeffding 1963, Theorem 1), we have

$$L^*(\beta) \leq L(\beta) \leq U(\gamma) \leq U^*(\gamma),$$

where $L^*(\beta)$ is the solution to

$$-2n(\overline{X}-r\overline{Y})^2 = \log\beta; \quad r < \overline{X}/\overline{Y}$$

and $1-U^*(\gamma)$ is the solution to

$$-2n(\overline{Y}-\overline{X}-r\overline{Y})^2 = \log\gamma; \quad r > \overline{X}/\overline{Y}.$$

Solving the above equations we get

$$L^*(\beta) = \overline{X}/\overline{Y} - [(-\log\beta)/2]^{1/2}/(\overline{Y}\sqrt{n}), \quad U^*(\gamma) = \overline{X}/\overline{Y} + [(-\log\gamma)/2]^{1/2}/(\overline{Y}\sqrt{n}).$$

Inequality (19) follows. $\square$

E-24-625 #2,3

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	11/30/95	Final Report Progress 8/1/94-7/31/95; 11/1/92-9/30/95

4. TITLE AND SUBTITLE
Progress and Final Report on AFOSR Project:
A Class of Methods for Analyzing Stochastic Systems

5. FUNDING NUMBERS
F49620-93-1-0043

6. AUTHOR(S)
Christos Alexopoulos

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0205

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/NM
Bldg. 410
Bolling AFB, DC 20332-6446

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)
Please see next sheet.

14. SUBJECT TERMS

15. NUMBER OF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT

This report summarizes the publications from our research on methods for analyzing stochastic systems. We studied three different system classes: (a) Probabilistic networks that model a variety of industrial and communications systems. These systems include data communications networks, voice communications networks, transportation networks, computer architectures, and electrical power systems. We corrected existing algorithms, derived the computational complexity of certain evaluations, and, based on new theoretical results, we proposed generalized algorithms that compute a performability measure by means of an iterative partition of the network state space. We also developed confidence intervals for Monte Carlo simulations tailored to the estimation of performability measures. (b) "Intelligent" Markovian networks where the processing of the units at the nodes and the routing of the units depend dynamically on the network congestion, and units can move concurrently. (c) Highly dependable systems with repairs. We have identified problems with existing simulation methods for estimating dependability measures and we are currently developing new methods that appear to be successful in a variety of large systems.

Progress Report
Final Report
AFOSR Grant F49620–93–1–0043
A Class of Methods for Analyzing Stochastic Systems

Christos Alexopoulos
School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332–0205

November 30, 1995

This document is the last progress report and final report for grant F49620–93–1–0043. My research has focused on three different areas, and has resulted in eight papers and two completed doctoral dissertations. Furthermore, one doctoral student is currently working under my supervision on problems that should be of great interest to Air Force laboratories and the airline industry.

The following sections describe my contributions in each research area during the last three years.

During the last year (8/1/94 to 7/31/95), I worked on papers 2, 5, 6, 7, 8 described in Sections 1 and 2, and I advised a doctoral student on the topic described in Section 3.

1 Probabilistic Networks

Probabilistic networks are used to model a variety of industrial and communications systems. These systems include data communications networks, voice communications networks, transportation networks, computer architectures, and electrical power systems. Stochastic networks are modeled by graphs in which each arc, and probably each node, is assigned a nonnegative random weight. The component weights have interpretations depending on the type of network under consideration. My research has focused on the evaluation of general *performability* measures and considered the following types of systems:

Flow Networks The nodes model distribution centers and the arcs represent the means of transmitting commodities between pairs. The nodes are classified into sources, demand nodes, and transshipment nodes. The weight on each arc and transshipment node represents a capacity that limits the total amount of commodity that can be transmitted. An arc may also be associated with a random cost per unit of transmitted commodity.

The following are typical measures of interest: (a) The probability that the demands can be satisfied; (b) The probability that a given set of links and nodes limits commodity transmission when the demands cannot be satisfied; (c) The expected amount of unsupplied flow when the demands cannot be satisfied; (d) The probability that the total cost for satisfying the demands does not exceed a specified value.

Transportation Networks The arcs represent sections of routes and the nodes represent intersections of routes. The weight of an arc represents its length or travel time. A list of interesting problems includes the computation of: (a) The distribution of the shortest path length from a source s to a destination t; (b) The probability that a given arc belongs to a shortest path.

Undirected Networks Networks with undirected arcs are often used for modeling communications systems or for solving a variety of problems. An example is a graph whose

arcs have random costs and the objective is the evaluation of the probability that the nodes can be connected via a spanning tree whose total cost does not exceed a given budget.

The majority of problems for computing performability measures for stochastic networks are *#P-hard.* This property has motivated the research for approximation methods (see [2] for a comprehensive review of the relative literature). One class of these methods attempts to compute bounds while another class focuses on Monte Carlo estimation methods.

My research has focused on a methodology that are based on iteration and, in short, evaluate a performability measure as follows: At each iteration, a subset of the system state space is partitioned into sets with known contribution to the measure, sets with zero contribution, and *undetermined* sets whose value is unknown. The method continues in the same fashion until no undetermined sets remain. The proposed methods have the following important properties:

- After each iteration, they produce lower and upper bounds that improve continuously.
- The bounds along with the remaining undetermined sets can be used for designing Monte Carlo sampling plans that (a) yield estimates with variance smaller by several orders of magnitude than the variance of the respective estimates produced by a crude Monte Carlo experiment with equal sample size and (b) take less time than the crude experiment.

The following is a summary on the papers in this area.

1. A Note on State-Space Decomposition Methods for Analyzing Stochastic Flow Networks by the PI. *IEEE Transactions on Reliability.* 44(2), 354–357, 1995.

Consider a flow network with single source s and single sink t with demand $d > 0$. Assume that the nodes do not restrict flow transmission and the arcs have finite random discrete capacities. This paper has two objectives: (1) It corrects errors in well-known algorithms by Doulliez and Jamoulle [1] for (a) computing the probability that the demand is satisfied (or network reliability), (b) the probability that an arc belongs to a minimum cut which limits the flow below d, and (c) the probability that a cut limits the flow below d; (2) It discusses the applicability of these procedures.

The D&J algorithms are frequently referenced or used by researchers in the areas of power and communication systems and appear to be very effective for the computation of the network reliability when the demand is close to the largest possible maximum flow value. Extensive testing is required before the D&J algorithms are disposed in favor of

alternative approaches. Such testing should compare the performance of existing methods in a variety of networks including grid networks and dense networks of various sizes.

2. State Space Partitioning Methods for Stochastic Shortest Path Problems by the PI. To appear in *Networks*.

This paper describes methods for computing measures related to shortest paths in networks with discrete random arc lengths. These measures include the probability that there exists a path with length not exceeding a specified value and the probability that a given path is shortest. The proposed methods are based on an iterative partition of the network state space and provide bounds that improve after each iteration and eventually become equal to the respective measure. These bounds can also be used for constructing simple variance reducing Monte Carlo sampling plans, making the proposed algorithms useful for large problems where exact algorithms are virtually impossible. The proposed approach differs from existing approaches in that it attempts to derive "optimal" partitions. The algorithms can be easily modified to compute performance characteristics of stochastic activity networks. Computational experience has been encouraging as we have been able to solve networks that have presented problems to existing methods.

3. State Space Decomposition Methods for Solving a Class of Stochastic Network Problems by Jacobson, J. A. PhD Dissertation, Georgia Institute of Technology, 1993.

This study focuses on state space partitioning techniques for computing measures related to the operation of stochastic systems. These methods iteratively decompose the system state space until the measure of interest has been determined. The information available in each iteration yields lower and upper bounds on this measure, and can be used to design efficient Monte Carlo estimation routines. We present here new theoretical results identifying strategies for significantly enhancing the performance of these algorithms. Using these results, we describe a generalized algorithm that can easily be tailored to address a variety of problems. We next use this algorithm to analyze two important models in the area of stochastic network optimization.

The first model concerns the probabilistic behavior of minimum spanning trees in graphs with discrete random arc weights. Specifically, we compute the probability distribution of the weight of a minimum spanning tree and the probability that a given arc is on a minimum spanning tree. Both of these problems are shown to be #P-hard but the *matroidal* structure of the minimum spanning tree problem gives rise to an impressive algorithm for computing the probability that an arc belongs to a minimum spanning tree.

The second model considers minimum cost flows in networks with discrete random arc costs and capacities. We consider the case of statistically independent costs and capacities for each arc as well as the case in which the cost and capacity of each arc

change simultaneously. In each case, we show that the evaluation of the distribution of the minimum cost flow for a fixed demand configuration is a #P-hard problem. Numerical examples are given throughout the thesis.

Overall, this thesis makes the following contributions:

- Advances the understanding of state space partitioning methods. In doing so, it makes these methods more accessible and draws strong conclusions about the performance of certain types of partitions.
- It proposes areas in which further gains can be made with regards to these powerful computational techniques.

We have written a lengthy paper that is going to be published in a special issue on reliability of an archival journal. A second paper is in the final processing stage.

4. Distribution-free Confidence Intervals for Conditional Probabilities and Ratios of Expectations by the PI. *Management Science*. 40(12), 1748–1763, 1994.

Many simulation experiments are concerned with the estimation of a ratio of two unknown means, the estimation of a conditional probability being an example. This paper proposes confidence intervals for the case in which the ratio is estimated by using independent, identically distributed random pairs with bounded and ordered components. Emphasis is given to the case in which each component can be expressed as the product of a Bernoulli and a bounded random variable. The proposed intervals result from distribution-free, Bernstein-type bounds on error probabilities, are valid for every sample size, and their asymptotic width decreases at the same rate as that of confidence intervals based on the central limit theorem. Experimental results show that the proposed intervals are conservative with superior coverage for small sample sizes (≤ 50). This superiority over "normal" confidence intervals makes them useful for Monte Carlo experiments for estimating performability measures of probabilistic networks.

5. Conservative Confidence Intervals for Multinomial Probabilities by the PI and A. F. Seila. To appear in *Operations Research Letters*.

Multinomial data are often produced as a result of survey sampling where questions may be answered by selecting one of a set of mutually exclusive choices. For example, suppose that a system has $k-1$ mutually exclusive failure modes and cell i represents the event that the system fails according to mode i in a specific time period. The event that the system does not fail is represented by cell k. A simulation run of n independent replications will produce multinomial data providing the number of replications in which the system did not fail, or failed according to each failure mode.

This paper proposes distribution-free confidence intervals for multinomial experiments. Below, we briefly discuss the single, but important, result of this paper. Let $p = (p_1, p_2, \ldots, p_k)$ denote the unknown cell probabilities and suppose that we draw n samples. Let $n = (n_1, n_2, \ldots, n_k)$ be the observed counts and denote the observed cell proportions by $\hat{p}_i = n_i/n, i = 1, \ldots, k$.

The proposed confidence intervals have the form

$$\hat{p}_i \pm t/\sqrt{n}, \quad i = 1, \ldots, k$$

with simultaneous confidence coefficient

$$\Pi(k, p; n, t) = P\left[\bigcap_{i=1}^{k} |\hat{p}_i - p_i| < t/\sqrt{n}\right] \geq 1 - \alpha, \quad \alpha \in (0, 1).$$

Our methodology is based on the bound

$$\Pi(k, p; n, t) \geq G(k, p; n, t) \geq 1 - 2 \sup_p G(k, p; n, t),$$

where

$$G(k, p; n, t) = \sum_{i=1}^{k} \exp\left\{-nt\left[\left(1 + \frac{p_i(1-p_i)\sqrt{n}}{t}\right)\ln\left(1 + \frac{p_i(1-p_i)\sqrt{n}}{t}\right) - 1\right]\right\},$$

and finds the smallest t such that

$$\sup_p G(k, p; n, t) = \alpha/2.$$

A lengthy proof shows that $G(k, p; n, t)$ is minimized when $p_1 = \cdots = p_m = 1/m$ for some $2 \leq m \leq k$ and $p_i = 0$ for $i > m$.

The following table summarizes our findings. The last column lists asymptotically valid "normal" confidence intervals from Fitzpatrick and Scott [4]. The entries in column 3 are valid for all $k \geq 3$. For example, when $n = 500$ the intervals $\hat{p}_i \pm 1.67/\sqrt{500} = \hat{p}_i \pm 0.075$ have joint coverage probability at least 0.95 regardless of the number of cells. The inflated width is consistent with expectations and seems a reasonable price to pay for robustness against the usual normality assumptions.

$1-\alpha$	n	t	*asymptotic normal*
0.90	50	1.53	$\hat{p}_i \pm 1.00/\sqrt{n}$
	100	1.48	
	200	1.44	
	500	1.41	
	1000	1.40	
	∞	1.36	
0.95	50	1.67	$\hat{p}_i \pm 1.13/\sqrt{n}$
	100	1.62	
	200	1.58	
	500	1.54	
	1000	1.53	
	∞	1.48	
0.95	50	1.99	$\hat{p}_i \pm 1.40/\sqrt{n}$
	100	1.92	
	200	1.87	
	500	1.82	
	1000	1.79	
	∞	1.73	

6. Minimal Connected Enclosures on an Embedded Planar Graph by the PI, J. S. Provan, H. D. Ratliff, and B. R. Stutzman. Submitted for publication, 1995.

The purpose of this paper is to develop algorithms for combining regions formed by embedded planar graphs. Planar graphs are used to represent many systems with transportation networks (e.g., roads, rivers, rail) being examples. There are a variety of sources including the U.S. government for such databases. In these networks, edges represent transportation links augmented with additional edges for natural boundaries (e.g., rivers), man-made boundaries (e.g., power lines), and political boundaries (e.g., county lines), and vertices are formed from the intersections of these elements. Our work is motivated by applications in the areas of network design, reliability, distribution and logistics, and geographic information systems.

We study five problems of finding minimal enclosures on a connected plane graph. The first three problems consider the identification of a shortest enclosing walk, cycle or trail surrounding a polygonal, simply connected obstacle on the plane. We propose polynomial algorithms that improve over existing algorithms. The last two problems consider the formation of minimal zones (sets of adjacent regions such that any pair of points in a zone can be connected by a non-zero width curve that lies entirely in the zone). Specifically, we assume that the regions of the graph have nonnegative weights and seek the formation of minimum weight zones containing a set of points or a set of

regions. We prove that the last two problems are NP-hard and transform them to Steiner arborescence/fixed-charge flow problems.

2 Markovian Network Processes

Markovian network processes have been used for describing the movement of parts and supplies in manufacturing and distribution systems as well as the movement of telephone calls and data packets in communications systems. The distinguishing feature of my research in this area with Richard Serfozo and Akram El-Tannir is the emphasis on the next generation of "intelligent" networks where the processing of the units at the nodes and the routing of the units depend dynamically on the state of the network, and units move concurrently (as with batch processing).

Most of the existing theory on Markovian network processes is for networks in which the units operate independently and move one-at-a-time, and their routes are independent. Our goal is to enhance the understanding of those complex networks by describing their stochastic behavior.

My two joint publications in this area are listed below.

7. A Multivariate Generalization of Markov Modulated Processes by the PI, A. El-Tannir, and R. F. Serfozo. Submitted for publication, 1995.

Markov modulated processes model queueing systems where the arrival and service rates vary according to a Markov process independently of the number of customers in the system. These processes, however, do not cover systems where the arrival and service rates depend on the number of customers present. An example is an $M/M/Y$ system where the number of servers $Y(t)$ at time t is a Markov process with rates that depend on the number of customers present.

This paper studies a family of multivariate Markov processes where transitions can take place simultaneously and the rate at which a set of components changes state depends on the state of the remaining components. This family covers a wide range of Markov processes including Markov modulated processes, Markovian queues with variable capacity, and standard network processes such as closed Jackson network processes. The paper makes the following two contributions: (a) It identifies processes whose stationary distributions have product form; (b) It presents approximations for stationary distributions. The main result proposes an approximation for a bivariate process (X, Y) based on an "auxiliary" process with "averaged" rates. When the component X has n states and the component Y has m states, the computation of the approximate distribution requires the solution of $m + 1$ subsystems each with dimensions $n \times n$ instead of solving an $(mn) \times (mn)$ system. We proceed by generalizing this result for multivariate processes, and conclude with additional approximations. We illustrate the proposed

techniques by analyzing the equilibrium behavior of several practical systems.

8. Partition-Balanced Markov Processes by the PI, A. El-Tannir, and R. F. Serfozo. Submitted for publication, 1995.

When can the stationary distribution of a Markov process be obtained by pasting together several stationary distributions that represent the process restricted to certain subspaces? This study describes a class of "partition-balanced" Markov processes that have this cut-and-paste or divide-and-conquer property. The importance of this property is that the problem of obtaining a stationary distribution on a large space (e.g., for networks) reduces to finding several stationary distributions on smaller subspaces, either by analytical means or simulations or by a combination of both.

The notion of partition-balance is a "macro-reversibility" property resembling the detailed balance property of reversible processes. We present several characterizations of partition-balance and identify subclasses of treelike, starlike and circular partition-balanced processes. A new circular birth-death process is used in the analysis. The results are illustrated by a queueing model with controlled service rate, a multi-type service system with blocking and a parallel-processing model. A few comments address partition-balance for non-Markovian processes.

3 Variance Reduction Methods for Simulating Highly Dependable Systems with Repairs

The development of methods for simulating highly dependable systems with repairs has been a popular research topic within the simulation and computer science communities during the last decade. Since failures in such systems are rare events, the estimation of system dependability measures such as the limiting (long run) unavailability and the mean time to failure require prohibitively long simulation runs.

A variety of papers (see [5] and [6]) have developed variance reducing techniques that use the importance sampling method. Specifically, those papers propose the combination of importance sampling with the regenerative method for estimating long run measures, and the combination of importance sampling with the conditional Monte Carlo method for estimating transient measures such as the average interval availability or the distribution of the interval availability.

Bruce Shultes, a doctoral student whose research has been funded by this grant during the last two years, attempted to improve on existing methods (e.g., *failure biasing* [5, 6] and *failure distance biasing* [3]) by using structural network information such as the state of a cut vector or a path vector. To our surprise, the existing methods failed to induce substantial variance reduction over the crude Monte Carlo method in networks

with complex structure. In several cases, they produced inflated variance estimates.

The following list contains our conclusions and results during the last two years.

- The existing methods are geared towards short paths to failure. Hence they have problems in systems with large cuts.
- The applicability of existing algorithms is limited to systems where each individual component is highly dependable. This limitation excludes systems whose dependability is due to redundancies at the component level.
- There exist near optimal importance sampling distributions that are non-stationary and appear to resolve the aforementioned problems.
- Several stationary importance sampling distributions that appear to perform better than existing methods have been identified.

Bruce Shultes will graduate by the Summer of 1996. The contribution of this grant to his academic achievements will be acknowledged in his dissertation as well as in the subsequent publications on this topic.

References

[1] Doulliez, P. and E. Jamoulle. Transportation networks with random arc capacities. *R.A.I.R.O.*, 3:45–60, 1972.

[2] Ball, M. O., C. J. Colbourn, and J. S. Provan. Network reliability. Chapter 11 in *Handbooks in Operations Research and Management Science: Network Models*, 7:673–762, 1995.

[3] Carrasco, J. A. Failure distance based simulation of repairable fault-tolerant systems. In *Proceedings of the 5th International Conference on Modeling Techniques and Tools for Computer Performance Evaluation*, 351–365, 1994. Elsevier Science Publishers B. V., Amsterdam.

[4] Fitzpatrick, S., and A. Scott. Quick simultaneous confidence intervals for multinomial proportions. *Journal of the American Statistical Association*, 82:875–878, 1987.

[5] Goyal, A., P. Shahabuddin, P. Heidelberger, V. F. Nicola, and P. W. Glynn. A unified framework for simulating Markovian models of highly dependable systems. *IEEE Transactions on Computers*, 41(1):36–51, 1992.

[6] Shahabuddin, P. Importance sampling for the simulation of highly reliable Markovian systems. *Management Science*, 40(3):333–352, 1994.